

Coming in from the cold

Science in the Arctic cries out for better coordination — perhaps modelled on what happens in Antarctica.

It is abundantly clear that the Arctic ice-cap, land and ocean play an important role in the climate dynamics of our entire planet. The region has long been one of intense interest to ecologists, atmospheric researchers, climate scientists and meteorologists. And it is now accepted that the Arctic — vaguely defined as the lands and oceans inside the Arctic Circle — is particularly vulnerable to global change.

Two forthcoming events — the G8 summit in St Petersburg, Russia, in July and next year's celebration of International Polar Year — provide vital opportunities for scientists to make the case for a more concerted, international effort to study this desolate yet fascinating region.

Sea-ice and glacier retreat, permafrost melting, carbon fluxes, vegetation changes, biodiversity and species adaptation to climate change are just some of the phenomena that make the region so interesting, and scientists are only just beginning to understand the relationships between them. Their efforts to do so can surely benefit the Sami, the Inuit and other indigenous people living in the region, whose delicate environments are under siege from pollution, ozone depletion and various human health problems. They can also help the rest of us, most obviously by contributing to our understanding of climate change.

The region's importance with regard to this issue was reiterated in the 2004 Arctic Climate Impact Assessment, a literature review that captured the frantic pace of Arctic warming over the past century and its effects on plants, animals and people. It also charted how changes there could ripple across the entire the planet (see page 146).

The assessment report, which drew considerable public attention, was the culmination of various research activities that were ignited half a century ago by the 1957–58 International Geophysical Year, which focused on polar research issues. The forthcoming International Polar Year, in 2007–08, will give polar researchers a chance to showcase their work and plan new approaches to further enhance our understanding of the Arctic.

Despite the stark findings of the 2004 climate assessment, the eight nations with territory north of the Arctic Circle — Russia, Canada, the United States, Denmark (on behalf of Greenland), Iceland,

Finland, Norway and Sweden — remain too passive in their approach to coordinating polar research. Their benign neglect has led to the gradual deterioration of parts of the network of meteorological stations in the Arctic (see page 133). Better baseline support for such monitoring would cost little, but would make a huge difference to Arctic researchers of all disciplines.

As part of its contribution to International Polar Year, Canada has allocated about Can\$90 million (US\$80 million) for this and related projects. Other countries, including the United States and Russia, have promised to improve Arctic and Antarctic monitoring. Scientists should check that these pledges are properly implemented.

By and large, scientists working in the Arctic are well connected with each other. But in contrast with Antarctica, where a 1959 international treaty obliges its signatories to collaborate in scientific research, there is no political framework for collaboration on Arctic research. The governments involved discuss regional issues through the Arctic Council, but this has no formal provision for cooperation in research.

"In contrast with Antarctica, there is no political framework for collaboration on Arctic research."

Contacts between Russia and the other nations with territory there improved after the end of the cold war. But many parts of Russia, which has by far the largest Arctic territory, are still not easily accessible to scientists, and there aren't enough truly international projects in the region.

Climate change will be on the agenda at the G8 summit, and the meeting provides an opportunity to improve the links between the large Arctic research communities of Russia, the United States and Canada. Leaders at the summit should commit to closer and more open collaboration on the model that has been tried and tested in the Antarctic.

Russia, which chairs the Arctic Council until the end of this year, should use its leadership to set up a council working group on pan-Arctic research that could address such issues as the availability of research permits and the maintenance of infrastructure for monitoring conditions in the region. ■

Special provision

Some research centres are more equal than others.

From a distance, it sounds like an event worth celebrating. At Westminster on 4 May, British government officials and scientists gathered to toast the Tyndall Centre for Climate Change Research. In its first five years, the Norwich-based centre has brought together social and natural scientists and produced

internationally respected work on the options that exist for responding to climate change. The gathering marked the award of another three years of support for the centre.

But behind the scenes, things aren't quite as rosy as they appear. One of Britain's more successful interdisciplinary research centres, the Tyndall centre is in fact facing a cut of some 15% in the real value of its income. Its research programmes will be reduced in scale and its PhD studentship programme abandoned. It also faces an uncertain future when current funding expires in 2009.

The Tyndall centre is a victim of two sets of circumstances that



are squeezing permanent research centres of its type. Until last year, the Treasury made special provision to protect the Tyndall centre's income. When that funding expired, the three research councils that oversee the centre, led by the Natural Environment Research Council (NERC), brought in outside experts to help decide what should happen next. But staff at the Tyndall feel that they lost out in this review because their work cuts across the expertise of the councils and the reviewers.

Additionally, all UK research agencies are under pressure to divert funding from permanent centres (such as the Tyndall) to the more flexible and efficient mechanism of individual investigator grants. In the past year, in various different circumstances, the National Institute for Medical Research (run by the Medical Research Council), the NERC's Centre for Ecology and Hydrology, the Tyndall centre and the John Innes Centre, a plant-science institute also based in Norwich, have all come under pressure from this general preference.

The emphasis on individual grants is generally a good thing. But it can be taken too far: unlike, say, France or the United States, Britain has already shut down its more inefficient government laboratories. It needs to retain some permanent research centres in order to support important government functions, such as managing public health and the environment.

The NERC would argue that the Tyndall centre has already done a large part of its job by helping to build interdisciplinary research

capacity. After eight years, it might make sense to fund further work through competitive grants; in Britain, these can be administered by more than one research council to support genuinely interdisciplinary projects. The NERC would also argue that its funding decisions have been reached after review of the Tyndall centre's work by independent experts.

However, the government badly needs the kind of research in which the Tyndall centre excels to help it make decisions about climate change. The centre's long-term survival would guarantee that this work keeps getting done. As well as maintaining the flow of useful reports, it would also provide a home for young researchers who wish to specialize in such interdisciplinary work but might struggle to find a career path in an atmospheric-physics or economics department.

Similar arguments apply at the UK Energy Research Centre, based in London, another cross-council project, which will see its own pot of dedicated funding expire in 2009. Its fate will largely rest on independent peer review. If it scores as highly as the Tyndall centre, the government should make special provision for both to guarantee their respective futures. ■

"The government badly needs the kind of research in which the Tyndall centre excels to help it make decisions about climate change."

Let the data flow

US legislation could fill a gap in drought research.

Compared with other types of natural disaster, drought rarely gets the public attention it merits. In east Africa, it is currently devastating entire nations — again. Even in wealthy nations it has a huge human and economic impact. Legislation currently before the US Congress acknowledges the potential role of science in drought preparation and mitigation.

The legislation would create a National Integrated Drought Information System (NIDIS) within the National Oceanic and Atmospheric Administration. The system would collect drought data, make forecasts and communicate information to the public. According to a bill introduced by Representative Ralph Hall (Republican, Texas), funding would start at \$12 million next year (a 50% increase over current federal drought funding), rising to \$18 million in 2012. Senator Ben Nelson (Democrat, Nebraska) has introduced a similar bill in the Senate.

NIDIS has enthusiastic support from western states, which bore the brunt of a US drought that ran from 1999 to 2004 and was among the worst for a century. As global warming kicks in, such events are likely to become more frequent and more severe.

Drought forecasting has made real advances in the past decade. The National Drought Mitigation Center at the University of Nebraska-Lincoln has taken the lead in producing an online US Drought Monitor, which tracks the scale and extent of drought nationwide.

Yet much more could be done. Today the relevant data come from a patchwork of federal, state, regional and local agencies. The

Department of Agriculture contributes snow-pack measurements, reservoir data come from the US Army Corps of Engineers and the Bureau of Reclamation, and the US Geological Survey (USGS) collects groundwater and stream-flow information. NIDIS could pull together, standardize and interpret the data they produce. Drought preparation would benefit from the same kind of focus that the National Hurricane Center in Florida currently provides.

NIDIS would also work to patch holes in the existing data network. Scientists involved in drought prediction could use better maps of soil moisture, for example. And experimental products such as the Vegetation Drought Response Index — designed by the USGS and other agencies to map the effects of drought at high resolution — could be developed more fully.

Funding for drought research may face some resistance from those suspicious that it is a ploy by western states to get some of the money that now goes to places hit by tornadoes and hurricanes. And all the research in the world won't help political leaders prepare for drought unless they come to grips with the contentious problem of water use in the West.

But it would be irresponsible not to learn more about the threat. Globally, there is no question that drought receives less attention than its dreadful consequences merit. Drought is slow, sometimes hard to define, and doesn't look spectacular on television. Senator Nelson has resorted to naming the recent drought in his state 'Drought David' in an attempt to raise awareness. Establishing NIDIS would be an even better idea. ■

"Drought preparation would benefit from the same kind of focus that the National Hurricane Center in Florida currently provides."

RESEARCH HIGHLIGHTS

Cut-price genes

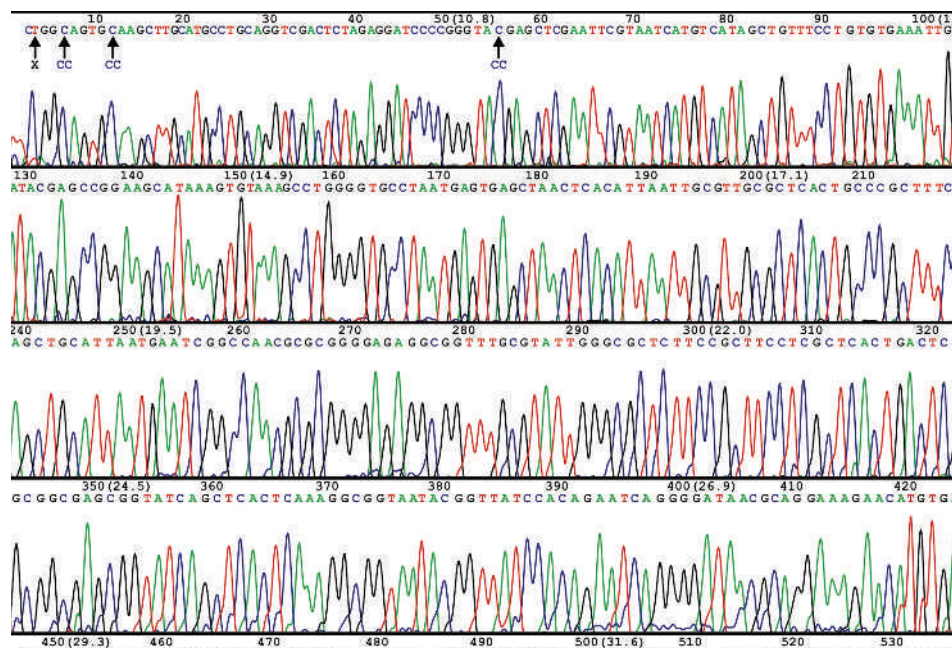
Proc. Natl Acad. Sci. USA

10.1073/pnas.0602476103 (2006)

It currently costs about US\$10 million to sequence 3 billion base pairs — roughly the size of the human genome — according to the US National Institutes of Health. New devices will help to cut the price. Richard Mathies of the University of California, Berkeley, and his co-workers present one option.

They have developed a 'lab on a chip' device that performs all three steps in an established DNA-sequencing technique, known as the Sanger method. The device should reduce costs by minimizing the amount of reagents needed and simplifying the process.

The researchers used the technique to sequence up to 556 continuous bases from a remarkably small sample of DNA — just 1 femtomole. The results (pictured) were 99% accurate.



NATL ACAD. SCI. USA

ARCHAEOLOGY

All that glitters

Archaeometry 48, 229–236 (2006)

Ceramicists have delighted in making lustrous, metallic glazes for far longer than we thought, according to findings from the Near East.

Metallic glazes are familiar in medieval pottery, but now Joris Dik of the Delft University of Technology in the Netherlands and his co-workers have found such material on fragments of a Levantine vessel at Deir 'Alla in Jordan, dating from the Late Bronze Age (1550–1200 BC). The vessel is decorated with designs in a greyish, sparkly glaze containing natural chromite. The metallic lustre comes from crystals of a calcium–magnesium–iron silicate called augite, the formation of which is induced by chromite.

CELL BIOLOGY

Nuclear history

Proc. Natl Acad. Sci. USA 103, 7077–7081 (2006)

Nuclear receptors — transcription factors that regulate genes in response to hormonal or metabolic signals — may have a much older heritage than previously thought.

Comparison of the sequences of different genomes has so far definitively identified nuclear receptors only in animals and sponges. But Didier Picard of the University of Geneva, Switzerland, and his colleagues have compared the predicted shapes of animal and fungus proteins to identify what may be a nuclear

receptor in the yeast *Saccharomyces cerevisiae*. If confirmed, this suggests that fungal and animal lineages evolved nuclear receptors before their divergence in the distant past.

EARTH SCIENCE

Going up with a splash

Geology 34, 349–352 (2006)

Researchers have discovered a new type of mantle plume in computer simulations of convection in Earth's molten layers.

Huw Davies of Cardiff University, UK, and Hans-Peter Bunge of the Ludwig Maximilians University in Munich, Germany, observed 'splash plumes' forming when a stream of downward-flowing cold matter, which could come from a sinking plate, met a sheet of hot mantle. The structure created looked like the splash of a water droplet — with splash plumes rising from

the rim of hot material thrown upwards.

It was previously thought that most plumes welled up from deep in Earth's mantle. Splash plumes may explain recent seismic images, which suggest that some plumes start in the middle of the mantle instead.

CIRCADIAN RHYTHM

CLOCK's surprising HAT

Cell 125, 497–508 (2006)

The body's master pacemaker — the circadian protein CLOCK — operates in the same way as enzymes known as HATs, says a study.

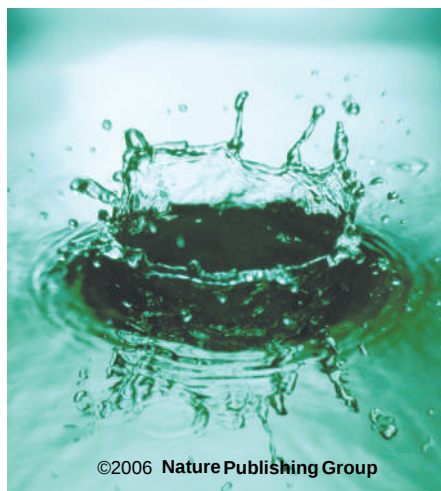
How CLOCK controls daily oscillations in the activity of a host of other circadian genes has been unclear. Paolo Sassone-Corsi at the the Institute of Genetics and Molecular and Cellular Biology in Strasbourg, France, noticed that portions of CLOCK look similar to enzymes known as histone acetyltransferases, or HATs. These enzymes tag an acetyl group on to histone proteins, around which the DNA in chromosomes is coiled. The researchers confirmed that CLOCK has HAT activity, and showed that cells in which the CLOCK protein lacked its HAT function lost their daily rhythms.

COSMOLOGY

Recycling space

Science doi:10.1126/science.1126231 (2006)

How did the cosmological constant (Λ) — a measure of the acceleration of the Universe's expansion — come to be close to zero, when



C. POTTON

theory predicts that it should be enormous?

Paul Steinhardt of Princeton University, New Jersey, and Neil Turok at Cambridge University, UK, say the solution might be that our Universe is cyclic. Such a universe would have expanded and contracted repeatedly on a trillion-year timescale, instead of having burst into existence only 13.7 billion years ago. This would have given Λ time to decay, avoiding the problem that cosmic expansion dilutes the Universe in the meantime — for matter is re-concentrated on each cycle.

EVOLUTION

Tropical mix

Proc. Natl Acad. Sci. USA

doi:10.1073/pnas.0510383103 (2006)

Is species richness in the tropics (pictured) fuelled by a faster rate of molecular evolution?

Shane Wright of the University of Auckland, New Zealand, and his colleagues compared 45 pairs of closely related plants from tropical and temperate climates. They found that the tropical species had more than double the average DNA mutation rate of their temperate cousins. They argue that increased mutation rate is the cause rather than effect of faster speciation, showing that tropical mutation rates were elevated even in a subset of genera whose temperate components were the more diverse.

The elevated metabolism associated with a warmer climate may be responsible for inducing mutations. However, the researchers could not completely rule out the possibility that genetic drift is simply more rapid in the smaller populations of the tropics.

CELL BIOLOGY

In three dimensions

Nature Meth. **3**, 369–375 (2006)

Traditional two-dimensional cell cultures, grown in the Petri dish, poorly mimic the natural organization of cells. But three-dimensional cultures have been difficult to control, prompting researchers in the United States to develop a fresh approach.

Sangeeta Bhatia of the Massachusetts Institute of Technology and her colleagues used an electric field to organize cells growing in liquid hydrogel into clusters of specific size and shape. The cells were fixed in place when the hydrogel solidified, forming sheets that could be stacked to build three-dimensional structures.

They used the system to show that the three-dimensional organization of chondrocytes, cells found in cartilage, influences the rate of synthesis of molecules important for cartilage function.



EVOLUTIONARY ECOLOGY

Live fast, die young

Proc. R. Soc. B doi:10.1098/rspb.2006.3544 (2006)

It has been observed that birds or mammals that develop very quickly as embryos tend to have short lives, and vice versa.

Robert Ricklefs of the University of Missouri in St Louis has used simple mathematical models to explore why. He suggests that the risks to an embryo of spending a long time in the egg or womb and the related cost of extended parenting are weighed against the advantages of having a long life in which to reproduce.

For example, sea birds, which have few predators, invest in a long development because their eggs are at a low risk and they can make the most of a long life. By contrast, animals that are likely to be picked off young come into the world in a hurry, and, because their chances of survival are low, are not built to last.

CANCER BIOLOGY

Off target

Proc. Natl Acad. Sci. USA **103**, 7444–7449 (2006)

Resistance to the archetypal 'targeted' anticancer drug imatinib (Gleevec) can occur when its target gene *BCR-ABL* mutates. But mutations in the tumour-suppressor gene *p53* can also reduce responsiveness to this drug, according to research.

Gleevec works by blocking the enzyme *BCR-ABL* kinase, which is encoded by its target gene. Scott Lowe, of Cold Spring Harbor Laboratory in New York State and his colleagues show that one effect of blocking this enzyme is to activate *p53*. They further show in a mouse model of leukaemia that inactivation of *p53* through mutation can reduce the anticancer effectiveness of Gleevec, even when its inhibition of *BCR-ABL* kinase is unaffected.

JOURNAL CLUB

Neill Alexander
University of Leeds, UK

A biomechanist wants to predict the energy cost of funny walks.

My main field of research is the mechanics of human and animal movement. I am not content to know how people and animals move: I also want to know why they move as they do.

Why do animals use particular, highly predictable stride frequencies, and exert particular patterns of force on the ground, when they walk or run at particular speeds? Why do people often lift things along curving trajectories, instead of directly vertically?

Questions such as these imply the further question, what would happen if the movement were done differently? But that is hard to discover experimentally. It is difficult enough to train people to perform funny walks, let alone animals. Therefore computer modelling often looks like the best way forward.

In many cases, the most promising hypothesis for why people and animals move as they do says that the chosen pattern of movement minimizes metabolic energy costs. To test this idea by modelling, we need to be able to predict the energy demands of muscles for possible patterns of movement that are not used.

I have tried to do this in studies of gait, with Alberto Minetti, now at the University of Milan, Italy, but have been concerned about the applicability of available data on muscle metabolism.

Now there is a paper that shows how muscular energy costs can be predicted with better accuracy (G. A. Lichtwark and A. M. Wilson *J. Exp. Biol.* **208**, 2831–2843; 2005). The paper tests the improved model against experimental data from muscles oscillating smoothly between contracting and stretching.

For the research that I want to see done, the approach will need to be validated for less-regular movements, but this paper takes a valuable step in the right direction.

NEWS

Arab state pours oil profits into science

How would you spend the profits from an oil well? That question was on the mind of about 200 Arab scientists who gathered in Doha, Qatar, late last month.

The country's head of state, Emir Hammad bin Khalifa Al-Thani, has devoted a chunk of Qatar's fossil-fuel profits to research, and the region's expatriate scientists were brought together to advise how best to invest it. Get it right, said attendees, and the money could help create an internationally respected science base for the Arab world.

"If it succeeds, it will change the whole region," says Hilal Lashuel, a Yemeni neuroscientist at the Swiss Federal Institute of Technology (ETH) in Lausanne. "Arabs will compete for the first time."

Qatar is already gaining scientists' attention in part because the country has recently transformed its university system. The tiny Gulf state, which has a population of less than 900,000, supports Western standards of living thanks to its substantial oil and gas fields. But those reserves are limited.

Hence Qatar's decision to jump on the knowledge-economy bandwagon and create Education City. The 2,500-acre campus on the outskirts of Doha hosts undergraduate teaching branches of several well-known US universities, such as Texas A&M. By replicating Western academic culture in the Gulf, the country's leaders hope to eventually attract 2,000 students annually to what could become the region's premier teaching facility. Several hundred students are already enrolled.

Education City has been bankrolled by the Qatar Foundation for Education, Science and Community Development. This is funded by an endowment from Al-Thani that officials say runs to billions of dollars. With this education centre up and running, the foundation is now turning its attention to applied research and the income-generating technologies that flow from it.

The emir has set aside profits from one of the country's oil wells for the purpose. Combined with a contribution from the foundation's endowment, Qatar will have a dedicated research fund worth hundreds of millions of dollars a year.

That may be small change in international terms, but Arab scientists say that the other components of Qatar's research vision show that the country means business. Foreign labs are being enticed to a science park on the Education City site, for example, by offers of state-of-the-art research

"Arabs will compete for the first time."

facilities and liberal intellectual-property agreements. All the labs have to do is provide the staff. It is a deal that has attracted interest from research institutions such as Imperial College London and Tokyo University.

Tempting talent

Tidu Maini, Imperial's pro-rector, says his university is working with the foundation on plans for a diabetes genome centre. The Gulf region has one of the highest rates of diabetes in the world, making it the ideal place to probe the genetics of the disease. Although plans are at an early stage, Maini says research groups could be set up this year at Imperial and the Hamad General Hospital in Qatar. The researchers would eventually move to the new genome centre, which Maini says could open in about three years. He would like every newborn child in Qatar,

and their parents, to be genotyped in order to build up a database that could be probed for diabetes studies.

Qatar is importing Western practices as well as personnel. The country's research fund, for example, will be administered by independent peer-review panels. These will be composed of researchers from home and abroad, and modelled on processes followed by organizations such as the US National Institutes of Health. Fathy Saoud, a parasitologist who sits on the Qatar foundation's board of directors, says that grant applications for biomedical, environmental and computing projects will be considered in about a year's time.

At last month's conference, held from 24 to 26 April, expatriate Arab researchers talked about how the Qatar project could boost science investment across the region. Researchers say that if the initiative takes off, it could force neighbouring countries to launch similar



Qatar's recently launched teaching centre, Education City, is helping to attract researchers from overseas.

M. MARION/WCMC-Q

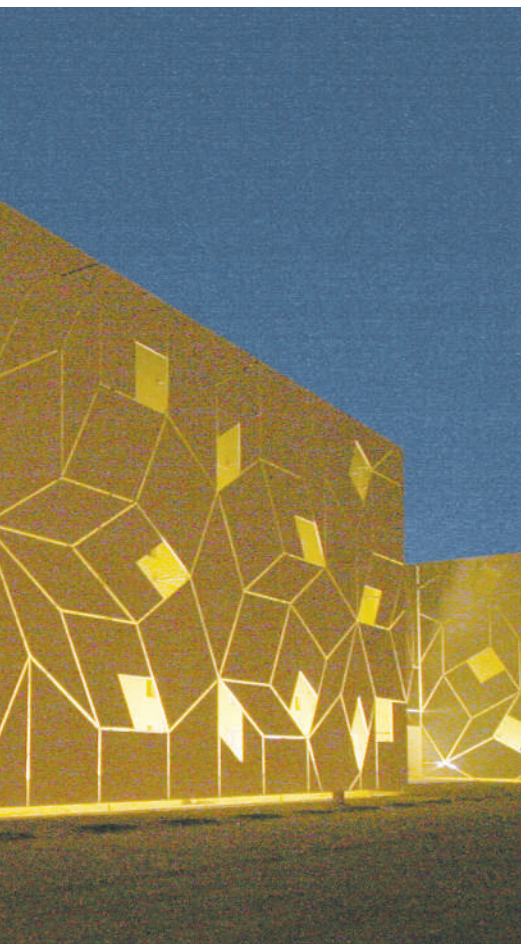


POPULAR PHYSICS MYTH IS ALL AT SEA

Does the ghostly Casimir effect really cause ships to attract each other?

www.nature.com/news

Arctic stations need human touch



projects and ultimately reverse the current exodus of bright students from the region to Europe and the United States.

"It's an idea whose time has come," says Abdelwahab El-Affendi, a political scientist at the University of Westminster in London who attended the Doha meeting. "There are large numbers of Arab expatriate scholars, who are good in their own areas, but their talents are not being used to advance research in the Arab world."

One potentially divisive issue concerns the involvement of the region's research powerhouse: Israel. Delegates at the meeting said that they were keen to collaborate with Israeli colleagues, but that they did not know whether that could happen before peace was reached with the Palestinians. "I don't think that Israelis can be involved at this juncture," says El-Affendi. "We have to wait for a genuine peace process."

Jim Giles

Arctic climate research is suffering as manned weather stations are being closed in Canada, Russia and the United States, some meteorologists complain.

Since 1990, around a quarter of the 500-odd manned meteorological stations in North America and Siberia have been shut in order to save money. Some have been replaced with automated monitoring stations, but other closures have completely halted data flow from those sites.

Among other variables, these stations measure winter precipitation and snow depth — key factors for assessing climatic and hydrological changes in the Arctic, where warming is more pronounced than at lower latitudes (see page 146).

"Each station less means we have one important grid point less for validating our models," says Konrad Steffen, an ice specialist at the University of Colorado in Boulder.

The reasons for the closures vary, but essentially come down to cost. A manned weather station can cost more than US\$100,000 a year to operate. Automated stations, around 3,000 of which are scattered across the Arctic, cost approximately \$30,000 a year, but can fall short in terms of data quality and reliability, many scientists say.

"It's just a terrible mess if there is no human around who can go out and check the equipment," says Jessie Cherry, a hydrologist at the University of Alaska in Fairbanks. "The main problems are mechanical failures, frozen gauges and under-catch of snowfall."

National weather services are required to report all weather observations to the World Meteorological

Organization (WMO), a UN agency based in Geneva, Switzerland. But the WMO cannot stop governments from closing stations.

In Canada's far north, a switch to automated measurements and the closure of several manned stations has reduced by 50% the number of stations with records suitable for monitoring snow cover.

But Thomas Nichols, director-general of weather and environmental monitoring at the government agency Environment Canada, argues that the country is still collecting plenty of data. Canada has installed 45 automated Arctic stations over the past four years; some

"It's just a terrible mess if there is no human around who can go out and check the equipment."

of these have been built at new locations, and others have replaced manned stations or ageing automated stations.

But hydrologists say data from the automated stations can be suspect. "It is difficult to believe what you are seeing," says Cherry, who has just completed a hydrologic study in Siberia based on the 60-year record of seven Russian weather stations (see doi:10.1038/news060403-9). "Given the number of possible biases out there, we just don't trust the raw gauge data."

Decreasing monitoring ability on the ground is also limiting the value of satellite-based snow and ice observations, which need to be calibrated against ground measurements.

The situation is similar in Antarctica. Russia and South

Africa have recently closed three manned stations there, which they plan to replace with automatic stations. Around 20 manned stations remain on the continent, along with 25 automatic ones.

Getting real-time data is a problem for the WMO's entire Global Climate Observing System (GCOS), says Phil Jones, a climate researcher at the University of East Anglia, UK. The GCOS collects data from about 1,000 terrestrial stations worldwide, 10% of which are in polar regions.

But each month, 30% of the stations don't send data, or send them late. "Data flow is alarmingly bad," says Jones. "It wouldn't take much effort to improve it, but some countries just don't seem to be concerned about data quality."

The lack of good data also troubles the Greenland research community. In northern Greenland, where glacier movement may be speeding up substantially, the scarcity of data is limiting scientists' ability to accurately model the phenomenon, says Eric Rignot, a glaciologist at NASA's Jet Propulsion Laboratory at the California Institute of Technology in Pasadena. "There are too many unobserved variables, such as glacier thickness close to the coast and meltwater movement," he says.

Scientists hope that the governments of Arctic countries will set up new manned monitoring facilities as a contribution to the International Polar Year, in 2007-2008.

The United States, Canada, Sweden and Russia have promised to support a number of projects aimed at improving Arctic monitoring (see page 127).

Quirin Schiermeier

ON THE RECORD

“Please post my seeds in a plain, unmarked envelope with no indication of contents to ensure smooth arrival.”

Horticulturalist Susan Davies gives instructions for delivering illegal rhododendron seeds to her New Zealand home. She was fined US\$3,200 for violating the nation's biosecurity act.

“Actually, King Tut has been flattered by the embalmers' work.”

Mummy expert Eduard Egarter Vigl reports on the state of the king's mummified penis, which was recently found in the sand around his body.

Sources: *Manawatu Standard*, *The Times*

SCORECARD

**Taxis**

The Pentagon launches a competition to build an autonomous vehicle that can navigate city streets.

**Jellyfish**

German scientists show that jellyfish perform one of the fastest cellular processes in nature: they eject their stinging cells in just 700 nanoseconds.

**Contraceptives**

The Vatican is reconsidering its rules on condom use — but any lifting of the ban would apply only to married couples with HIV.

OVERHYPED

Coma drama

Hollywood scriptwriters love putting characters into comas, but irate neurologists say they are fudging the facts. The most common error in 30 movies studied was the suggestion that coma patients keep their toned bodies and perfect tans over a period of years. The cinema also glosses over details of comatose life such as incontinence, feeding tubes and respirators. Most significantly, the authors point out, there is no evidence that patients waking from a coma immediately seek revenge.

Source: *Wijdicks, E. F. M. & Wijdicks, C. A. Neurology* **66**, 1300–1303 (2006).

GETTY



“They better damn well like it”

Outspoken: Mike Griffin on the NASA budget

NASA head Mike Griffin was blunter than usual last week, as he defended his scaling back of the agency's science programme. Space scientists have responded angrily to the cutbacks (see *Nature* **439**, 768–769; 2006), but Griffin insisted to two key advisory groups — the Space Studies Board and the science subcommittee of the NASA Advisory Council — that the science programme is still healthy. He made it clear, however, that the White House's plan to send astronauts to the Moon and Mars takes priority over increased science funding. And although he is willing to rethink some specifics, he reminded scientists “to be respectful of the political and budgetary constraints we face”.

Griffin's stance on...**Whether budget cuts might be reversed**

We're willing to reconsider, but reconsideration should be based on community input, not the loudest voice, the longest e-mail or who can use the most capitals.

The outcry over cutting research grants

The community doesn't care if we fly missions; they want money for universities. I find that, to be honest with you, appalling.

A law requiring NASA to try to rescue the Hubble telescope — even though such a mission would take hundreds of millions of dollars from other science projects

I hope the astronomy community likes the decision they lobbied for. They better damn well like it, because they got it.

Cutbacks to life-sciences research aboard the International Space Station

What is the point of funding life-sciences research when I can't put people into space? I need the budget I have to recreate abilities that we once had to fly [beyond Earth orbit], that we don't have any more. It's a sequencing problem.

Deep cuts to NASA astrobiology

I did think astrobiology was less important than traditional space science. It had less intrinsic subject matter to it, and was less advanced. If the community rises up and says it should be funded, we'll rethink it.

Opportunities for science on the Moon

I have to draw the line when people say “I'm not interested in the Moon. I would rather put the money into studying the physics of the tropopause.”

OK, great. Glad you have an opinion; everybody gets one. But the people who run the country have decided that we are in fact going to the Moon. It's a question of what scientists would like to do with that.

The importance of finishing the International Space Station

As administrator, I inherit a situation not of my liking. But other nations have spent a very significant part of their own discretionary space funding supporting our agenda. They built their hardware, and they want to see it flown. I want us to honour this commitment.

Tony Reichhardt

Are rich nations up for drug reform?

The global debate over the high cost of drugs reaches Geneva this month, when the governing body of the World Health Organization (WHO) will consider two proposals aimed at reshaping the forces that drive medical research and development.

The proposals direct the WHO to act on two controversial issues: the lack of affordable, effective medicines for people in poor countries, and the international intellectual-property system that governs the distribution of medical treatments.

Developed countries — especially the United States — have in the past strongly opposed such interventions, perceiving them as a threat to the drug industry. But this is the first time the issues have been addressed by the WHO, and advocates for reform are optimistic. Many believe that recent trends — such as the massive and growing cost of health care in the United States, the scarcity of treatments for avian influenza, and the lack of biodefence countermeasures against diseases such as anthrax — are persuading the Western world that the current system does not meet all public-health needs.

The first proposal to come before the World Health Assembly, the WHO's governing body, is raised

by Kenya and Brazil, and calls on the agency to consider measures that would boost the development of new drugs. The reform movement stems from a growing concern about the lack of treatments, vaccines and diagnostic products for diseases that affect developing countries. For instance, one study found that only 1% of drugs brought to market between 1975 and 1999 targeted tuberculosis and tropical diseases such as malaria

"An increasing percentage of new drugs are slight variations of already-approved medications."

(P. Trouiller *et al. Lancet* 359, 2188–2194; 2002). The other 99% targeted disorders more relevant to the developed world, such as cancer and heart disease.

But the lack of new treatments is now affecting richer countries too. Other studies have found, for example, that an increasing percentage of new drugs are slight variations of already-approved medications. "There's possibly a recognition that something's going to have to change," says Tim Hubbard, a genome biologist at the Wellcome Trust Sanger Institute in Cambridge, UK, who has become heavily involved in the issue. "Everyone can see

that spending on research and development is going up, but the number of new drugs coming to market is going down."

The proposal has been endorsed by a long list of consumer groups, activists and scientists, including Hubbard. It doesn't endorse specific reforms, although advocates have suggested many options, from an international fund for research into neglected diseases, to a cash prize for developers of innovative drugs.

The important thing, they say, is that the WHO commits to tackling the issue. "It's been an uphill struggle to get the WHO interested in this question, and if you look at the enormous health needs out there you think, how the hell can this be the case?" says Ellen 't Hoen of Geneva-based Médecins Sans Frontières.

The second proposal asks the WHO to act on the issue of whether patent laws restrict access to essential medicines. It follows a report commissioned by the WHO, released on 3 April, which concluded that the patent system is not working to drive innovation in drugs needed by poor countries. It makes 50 recommendations for action.

The proposals will be heard when the assembly meets on 22–27 May. Advocates are encouraged that rich nations have not yet taken up a strong position against them. ■
Erika Check

There is growing concern about the lack of treatments and vaccines for diseases that affect developing countries.

GETTY IMAGES



SPECIAL REPORT



O. KISSNER/AFP/GETTY IMAGES

Avian flu and the New World

The H5N1 avian influenza virus has not yet reached North and South America. What will happen when it does? **Declan Butler** and **Jacqueline Ruttimann** investigate.

The H5N1 virus once seemed a problem that was lurking in someone else's backyard farm. But since last summer, the lethal avian influenza virus has surged out of southeast Asia into Europe, the Middle East and Africa. Fifty-one countries — 36 this year alone — have now experienced outbreaks.

The New World is as yet untouched, but many experts consider it inevitable that the virus will reach North and South America. This area is home to the world's largest poultry exporters, Brazil and the United States. And although government and industry officials say they are well prepared for the arrival of H5N1, others argue that they have yet to take on board the full extent of the challenge.

"When H5N1 arrives," says Mark Cackler, the World Bank's agriculture manager for

Latin America and the Caribbean, "it will concentrate minds wonderfully."

Trade, smuggling and migratory birds are all potential routes for H5N1 to reach the Americas (see page 138). How it arrives is ultimately less important than stopping a subsequent spread of the disease, says Juan Lubroth of the UN Food and Agriculture Organization. And here there is no debate as to what causes spread, he adds: it's not migratory birds, but "human actions, trade and husbandry practices".

Experience in Asia and Europe provides best practices on controlling avian flu, says Jim Butler, director-general of the Inter-American Institute for Cooperation on Agriculture in San Jose, Costa Rica. "We don't need to reinvent everything." The task for the Americas, he says, is more about "building consciousness"

of the risks and existing best practices.

Such practices include setting up surveillance and diagnostic capacities that can quickly detect any initial outbreak. But despite all best efforts, some virus will always go undetected, so countries must also have policies in place to make their farming systems more resilient.

Such policies can be as simple as requiring the poultry industry to disinfect all material and people moving in and out of farms. Another is to reorganize national poultry industries into regional 'cells' that operate independently of one another, so that if one goes down with H5N1, the others do not.

France prepared by using both approaches, and so far the H5N1 found there has been stopped in its tracks. Turkey did neither well, ►

H5N1: INTO THE AMERICAS

Trade

Every day, millions of live poultry are moved around the world by ground, air and sea transport, which potentially could carry H5N1 to fresh areas.

Replenishing flocks at huge industrial farms is now a global business: commercial hatcheries export billions of hatching eggs and one-day-old chicks. Egypt, for instance, exported some 180 million chicks annually before the H5N1 virus was detected there and exports were halted. Many experts see trade as the major cause of the spread of avian flu, through infected bird droppings on shells, crates and other surfaces.

In the United States, imported live birds

must undergo a 30-day quarantine and tests for avian flu, says Madelaine Fletcher, a spokeswoman for the US Department of Agriculture. Most day-old chicks reaching the country are from Canada or the United Kingdom, and their crates are either disinfected or destroyed, she says.

Hon Ip, a virologist at the US National Wildlife Health Center in Madison, Wisconsin, says it is hard to assess how well quarantine works, as the United States has so far avoided an outbreak of H5N1. But an outbreak of contagious Newcastle disease in California in 2002 shows that quarantine is not always watertight, he says.

A country is generally banned from exporting poultry when it officially reports an outbreak of H5N1 in commercial flocks. But H5N1 can be present before a report is made. Moreover, although H5N1 has occurred in 13 countries in the European Union (EU), the United States bans imports only from regions involved, not from entire countries.

Most US imports from Europe are pet birds from Belgium and poultry products from France, but the country also imports bird skins and feathers from 16 EU nations. In 2005, the United States imported more than 16.8 million day-old chicks and other live poultry as well as 16.8 million hatching eggs.

Smuggling

On 5 September 2005, customs inspectors at the US Department of Agriculture (USDA) found an illegal shipment of 98,400 chicken eggs at a port in California. During October and November, they intercepted nearly 75 tonnes of poultry smuggled in from Asia. Last month, a man from Nigeria — a country with H5N1 — was stopped at Miami airport after dogs detected dead birds, ostensibly brought in for a religious rite, in his luggage.

How much smuggling goes undetected is anyone's guess, but the illegal trade in birds and poultry products is thought to be third only to narcotics and arms in value. With

trade bans imposed on countries with H5N1, the risk of smuggling has gone up, says Ip. "It has become more likely that illegally imported birds could come from an affected country," he says.

Inspections of markets and restaurants in major US cities regularly turn up illegal imports of poultry from Asia and elsewhere, admits one USDA official.

On the front line, 122 inspectors from the US Fish and Wildlife Service work across 35 ports looking for another target: exotic birds. The risk is real. In October 2004, a Thai man was caught at a Belgian airport as

he tried to smuggle in two eagles that were later found to have H5N1.

Most smuggled birds are parrots, macaws and their relatives, or songbirds, says Ip. They come mainly from H5N1-free Australia, South America and Mexico. But traffic is brought in from many other countries. The black-capped lory (*Lorius lory*) from Indonesia, an H5N1-infected country, is a highly prized bird that is frequently targeted by smugglers, he notes.

"An infected bird in one part of the world can be in North America within 24 hours, whether that bird is fit to fly or not," says Ip.

Migration

Migratory birds could bring H5N1 to the Americas as early as this month — to Alaska, a short hop across the Bering Strait from Russia. Some 6.6 million birds will arrive in the state by the end of May, travelling up a flyway from Siberia, China and southeast Asia. The fear is that these birds could infect American migrants, which would then carry the disease back down the length of the Americas in the autumn.

To prepare for this, US agencies are already catching birds for testing — the start of a procedure that will examine 75,000–100,000 swabs from the birds for avian flu. "It's the largest and most complex

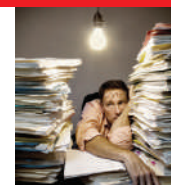
migratory bird capture and sampling programme in the history of Alaska," says Dirk Derksen of the Alaska Science Center in Anchorage, one of the project's main coordinators. Birds will also be monitored along the major flyways running down through North America, particularly along the West Coast.

H5N1 could also arrive from the east: birds from West Africa and Europe fly over Greenland into northern Canada. But the smaller number of birds and species, and the longer flight distances, make this a less likely route for H5N1, says Derksen, and so sampling is not taking place there.

Biologists have identified the relative risk of each species of Alaskan migrant. Top of the list is the northern pintail (*Anas acuta*), which breeds widely across North America and Eurasia. It is among the most highly infected of waterbirds, with one in ten birds carrying some flu virus (B. Olsen *et al. Science* **312**, 384–388; 2006).

But whether migratory birds can carry H5N1 over long distances remains controversial. The first outbreak of H5N1 in Africa — in Nigeria — was widely attributed to migratory birds. But many now see the imports of day-old chicks as a more plausible cause.

D.B.



PATENT OFFICERS CRACK UNDER PRESSURE

European workers strike against system to boost efficiency.

www.nature.com/news

GETTY

Continued from page 137

and has suffered persistent outbreaks. The well-developed animal health services in the United States and Canada have implemented similar measures to those in France, and are widely expected to quickly contain any outbreak of H5N1.

Unlike regions of southeast Asia and Africa, where poultry live outdoors and mingle freely with each other and with wild birds, US flocks are highly concentrated in industrial farms, where birds are raised indoors, isolated from external contaminants. One potential weakness is the popular live markets in California and the northeastern states. In the northeast, these markets have sustained an avian flu virus of low pathogenicity, H7N2, since 1994. In the past, as many as 60% of birds sampled have carried flu viruses, although control measures have reduced the levels.

Such live markets "continue to be a major source of avian influenza viruses and a risk for introduction to commercial poultry operations". This is what David Swayne and David Suarez of the US agriculture department's main poultry research laboratory in Athens, Georgia, told an influential 2004 workshop on pandemic planning that was organized by the US National Academies.

South America is also preparing well, experts say, as it has decades of experience of handling other highly contagious diseases, such as foot-and-mouth disease. Lubroth gives the continent good marks for "high awareness favouring quick detection and response".

Brazil's poultry industry is both larger and more complex than that in the United States. Backyard farms are common in the poor northeastern part of the country, and small-holdings are scattered throughout urban areas and along the Amazon. But its core production remains in large commercial farms, which are well aware of the economic risks of H5N1, says Lubroth. "If H5N1 entered," he predicts, "they would stamp it out quickly to preserve the main industry."

That view is overly optimistic, argues Cackler, who sits on the avian-flu taskforce for the World Bank. Many South American countries, he says, have only recently woken up to H5N1 as something they might need to worry about. He recalls speaking to the agriculture minister of "one to-be-unnamed country" last November, who "just didn't get it; H5N1 was nothing to worry about". Cackler met him again in January, after H5N1 had ripped across Europe. "This time, the guy got it." ■

State's flu response raises concern

Just hours earlier, crowds had thronged past rows of squawking chickens, ducks and geese at a live-bird market in Camden County, New Jersey. But late last month, inspectors shut down the bustling market, ordering its complete disinfection after discovering an H5 avian influenza virus.

In the end, the virus turned out to be a strain that was not very harmful, but the event sheds light on what might happen if H5N1 is detected in the United States.

The country has weathered three major outbreaks of highly pathogenic bird flu before (see 'Past US outbreaks'). 'Low-pathogenic' bird flu, which kills few infected birds, occurs far more regularly. In the latest case, New Jersey's agriculture department made a public announcement about the discovery of an avian-flu strain — but it left out salient details.

The announcement on 28 April did not mention when or specifically where the infection was detected, saying only that preliminary tests had marked it as negative for the neuraminidase protein N1. The statement did not mention the haemagglutinin protein; *Nature* learned later that the state had a faint positive for H5, which can occur in both high- and low-pathogenic strains. The first samples were tested on 21 April.

Later confirmatory tests by the US Department of Agriculture (USDA) laboratory



Discovery of avian flu led to the closure of a New Jersey live-poultry market.

AP PHOTO/J. SZYMASZEK

in Ames, Iowa, failed because technicians there could not grow the virus. In the meantime, other birds in the market had been killed and disposed of. The market was later reopened.

If a low-pathogenic strain of bird flu is discovered, then individual states, not the federal government, are responsible for alerting the public — and officials say this all went as planned. "The timeline was exactly as it should be," claims Andrea Morgan, veterinary administrator for the USDA Animal and Plant Health Inspection Service. "The response that New Jersey launched was appropriate."

But it was misleading, critics

argue. Jody Lanard, a risk-communication specialist based in Princeton, New Jersey, has worked as a special adviser on pandemic communication to the director-general of the World Health Organization. She notes that the state's two press releases omitted the fact that the strain was H5, focusing instead on the fact that it was not N1.

"They are afraid the public will hear H5 and go nuts — a case of official panic about panic," she says. "If they really think the public is that fragile, they might be tempted to hold back lots of preliminary information, and delay issuing material when it really matters."

Karen Eggert, a spokeswoman for the USDA, says the department is still working out how and when it would alert the public to outbreaks of highly pathogenic strains such as H5N1. In the interests of openness, officials are considering announcing it immediately after the first confirmatory molecular tests. ■
Jacqueline Ruttimann

Past US outbreaks

Three highly pathogenic strains of avian influenza have already swept through the US poultry industry.

1924: An H7 strain of bird flu, known as the fowl plague, raced through live-bird markets in the northeastern United States.

1983: Another outbreak in the northeast — this time of H5N2 — culminated in the destruction of 17 million chickens, turkeys and guinea fowl.

2004: H5N2 struck again but was rapidly contained at a single poultry farm in Texas. Only 6,600 birds had to be killed.

Senators want public access for all research papers

Two senators have introduced a bill that would require US researchers with public funding to make their papers freely available within six months of publication.

Under the proposed law, federal agencies with annual research budgets of more than \$100 million would be required to put all research findings in a digital, publicly available repository. The requirement would not apply to unpublished works, lab notes or content protected under patent law. The bipartisan bill was introduced on 2 May by senators John Cornyn and Joe Lieberman.

The plan comes nearly a year after the National Institutes of Health unveiled a voluntary scheme to make its research available within 12 months of publication (see *Nature* 433, 561; 2005). But fewer than 4% of eligible articles have been added so far.

Japan creates prize for work on African diseases

A major medical research prize, financially comparable to the Nobels, will be awarded by Japan starting in 2008.

The prize will go towards research and clinical work on diseases prevalent in Africa, Prime Minister Junichiro Koizumi announced in Ghana on 2 May.

The selection criteria and cash value have not yet been decided, but it will be named after bacteriologist Hideyo Noguchi, who currently graces the ¥1,000 (US\$9) bill. Noguchi died of yellow fever in 1928 in Ghana's capital — he was investigating the viral fever, which he thought might be bacterial in origin.

The prize follows Japan's promise to double aid to Africa between 2004 and 2007.

European grant ensures free data for biologists

The European Bioinformatics Institute (EBI), based near Cambridge, UK, has launched a project called FELICS (Free European Life-science Information and Computational Services) thanks to a €16.7-million (US\$21-million) grant from the European Union.

The EBI is Europe's largest curator of biological information, including genome sequences and protein structures. FELICS

High-level merger starts tracking galaxies

An array of 15 telescopes has begun probing galaxies near and far after being set up in the California mountains, near the Nevada border.

On 15 April, scientific operations began at the Combined Array for Research in Millimeter-wave Astronomy (pictured). It is a combination of six 10.4-metre telescopes from the Owens Valley Radio Observatory and nine 6-metre telescopes from the Berkeley-Illinois-Maryland Association array.

The telescopes were relocated, at a cost of \$15 million, to a site more than 2,400 metres



above sea level — to improve visibility. They will probe radio emissions from galaxies, molecular clouds, newborn stars and more.

R. PLAMBECK

should give researchers unrestricted access to important data.

The EBI has lurched from one financial crisis to the next, ever since its foundation in 1994. A previous EU grant ran out last July. But the new one, which runs for five years, “will allow us to meet the evolving needs of biologists in Europe and around the world,” says the EBI’s Cath Brooksbank.

The European Patent Office will provide FELICS with information about newly patented chemicals. The Swiss Institute of Bioinformatics is also a partner, as is the University of Cologne with its important enzyme database, BRENDA.

Atmospheric report rebuffs warming sceptics

The first report to emerge from the US Climate Change Science Program, concerning temperature trends in the lower atmosphere, concludes that “the observed patterns of change over the past 50 years cannot be explained by natural processes alone”.

The report, published on 2 May, also reaffirms that there is “no fundamental inconsistency” between warming in the troposphere, the lowest layer of the



Temperatures in the lower atmosphere are consistent with global warming.

atmosphere, and the predictions of climate models. Observations that had suggested a discrepancy were used by sceptics to question the accuracy of climate models. But scientists involved in compiling the report attribute the problems to experimental uncertainties and errors in data analysis — they published papers on the subject last year (see *Nature* 436, 896; 2005).

The government programme, which was set up in 2002, plans to issue 21 reports summarizing the state of climate research.

United States gives Iraq a virtual science library

The National Academies has unveiled a virtual library that gives Iraqi researchers access to thousands of scientific journals.

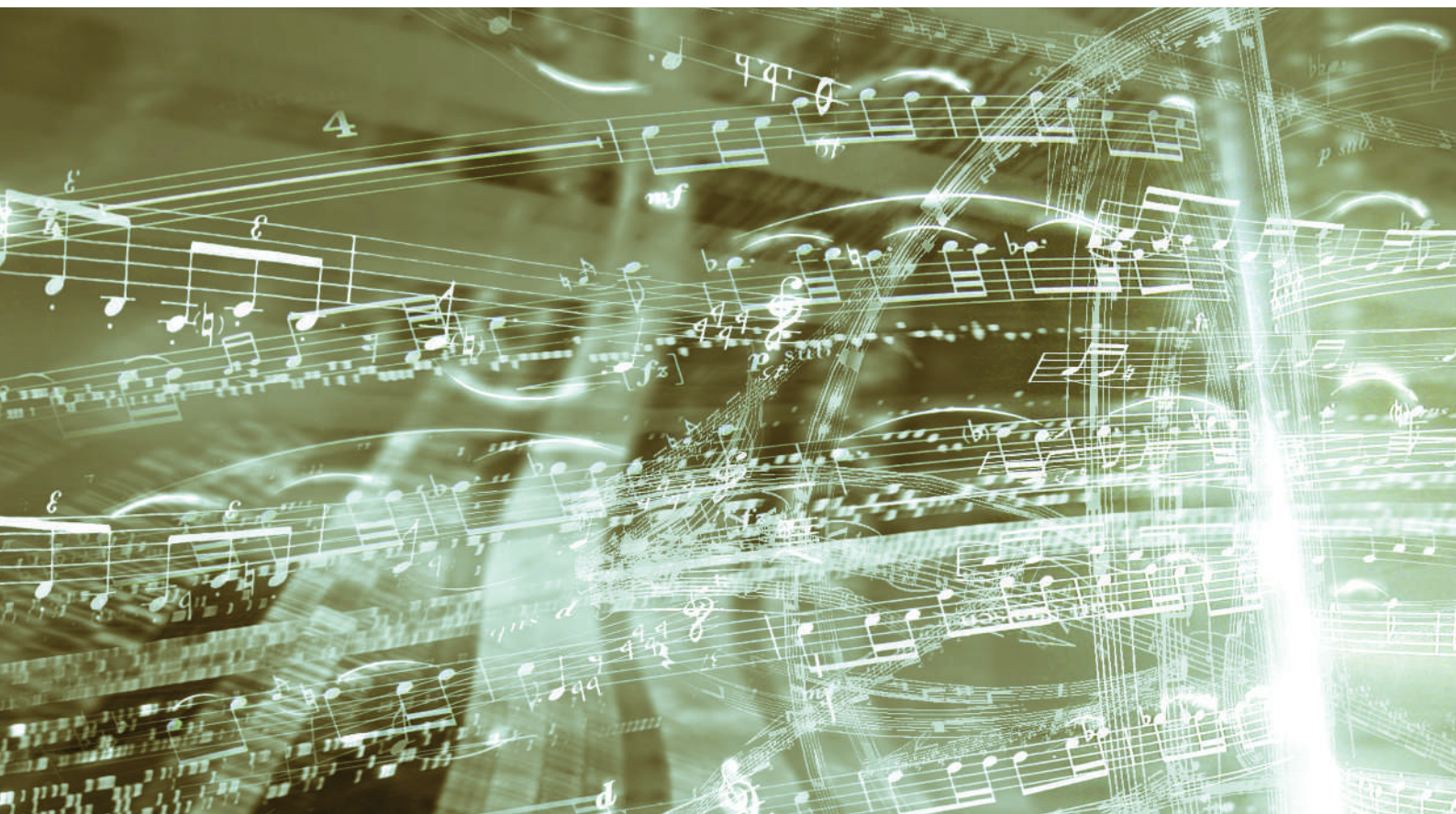
The Iraqi Virtual Science Library — a collaborative effort between the academies, the US government and publishers — will give scientists and engineers at approved Iraqi institutions access to more than 17,000 professional journals. The library also includes online courses and announcements of funding opportunities.

Samir Shakir Mahmud Al-Sumaydi, the Iraqi ambassador to the United States, called the new library “extremely significant” for his country’s struggling scientific and technical community. “Iraqi academics and scientists in particular are being targeted by terrorists,” he says. “The launching of this library sends a much-needed message of hope, support and solidarity.”

► <https://ivsl.org>

Correction

The article ‘Can super-antibody drugs be tamed?’ (*Nature* 440, 855–856; 2006) contained two errors. Steve Anderton researches autoimmune disease, not the use of antibodies to overcome cancer. And he did not state that the CD25 receptor may be a safer target for antibody therapies than CD28; instead, he suggested that CD25 may be of some use as a marker for regulatory T cells, and that other costimulatory molecules, such as CD154, may represent safer targets for immunomodulatory antibodies. *Nature* apologizes for any confusion.



UNFINISHED SYMPHONY

To correctly 'play' the DNA score in our genome, cells must read another notation that overlays it — the epigenetic code. A global effort to decode it is now in the making, reports **Jane Qiu**.

Manuel Esteller's phone did not stop ringing for weeks. It was summer 2005, and he and his team at the Spanish National Cancer Centre in Madrid had just published a study comparing the activity of DNA in identical twins. The anxious callers were invariably twins whose sibling had developed a serious disease such as cancer or diabetes. Could the study help predict whether they too would succumb, they asked. Did the identical DNA sequence they shared with their afflicted twin mean they had the same genetic predisposition to illness?

Surprisingly, the answer to the second question is 'not necessarily'. Researchers have known for years that, despite their common genes, identical twins can have very different physical constitutions and develop different diseases. The traditional explanation for this is that our environment somehow interacts with our genes to produce our physical attributes, or phenotype, but no one knew exactly how.

The study by Esteller and his team¹ showed

that the missing link between nature and nurture could lie in a phenomenon known as epigenetics: a cryptic chemical and physical code written over our genome's DNA sequence. The term 'epigenetics' was first coined in the 1940s by British embryologist and geneticist Conrad Waddington, to describe "the interactions of genes with their environment, which bring the phenotype into being"². The term now refers to the extra layers of instructions that influence gene activity without altering the DNA sequence.

"Deciphering the epigenetic code will illuminate some of the most profound questions in biology."

— Stephan Beck

on gene activity: the number of genes that differ in activity between 50-year-old twins was more than three times that in pairs aged 3. "So we are more than our genes," says Esteller. "Not only is the DNA sequence important but

also how gene activity is regulated in response to environment. This might explain why many identical twins have different susceptibility to disease."

Spot the difference

As well as offering answers to identical twins, deciphering this epigenetic code promises to dramatically alter our understanding of disease in the wider population (see 'Tagged for disease', overleaf). Many cancers might be triggered by epigenetic faults, for example. It should also fill some big gaps in our grasp of how the environment affects a creature's constitution — epigenetic changes explain how simply altering the diet of a pregnant mouse, for example, can completely change the coat colour of her pups³, or even alter their response to stress⁴.

"It will illuminate some of the most profound questions in biology," says Stephan Beck, an immunologist at the Wellcome Trust Sanger Institute, Cambridge, UK, who worked on the Human Genome Project. How a given cell executes its unique genomic programme in time and space could shed fresh light not only on development and disease but also

on what makes us human, he says.

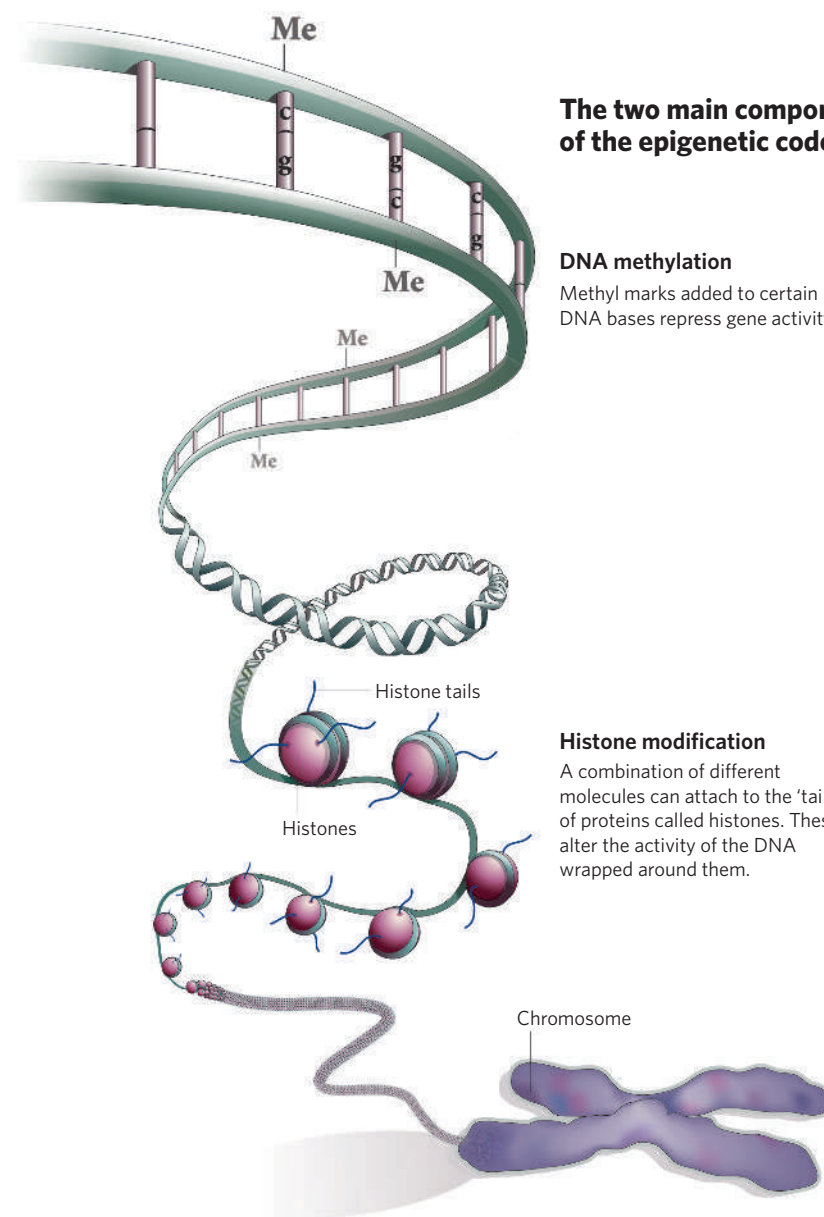
The complete epigenetic code of our genome, its 'epigenome' has increasingly been the focus of research over the past decade, and scientists are now embarking on an ambitious attempt to crack it. The International Human Epigenome Project, or IHEP, first suggested by Beck and colleagues in 1999, is the logical next step after the Human Genome Project, which published the draft sequence of the human genome's 3 billion DNA letters in 2001. But the IHEP faces daunting challenges. The sequence of the human genome is the same in all our cells, whereas the epigenome differs from tissue to tissue, and changes in response to the cell's environment. Can researchers really hope to pin down this vast, complex and ever-changing code in a meaningful way?

Clever packaging

If the DNA sequence of the genome is like the musical score of a symphony, then the epigenome is like the key signatures, phrasing and dynamics that show how the notes of the melody should be played. Epigenetic control of gene expression occurs in two main ways: either the DNA itself is chemically altered, or the proteins that package DNA into chromatin (the main component of chromosomes), are modified. These proteins, called histones, determine whether the chromatin is tightly packed, in which case gene expression is shut down (or silenced), or relaxed, in which case gene expression is active.

The first kind of alteration takes the form of methyl groups added to the DNA — frequently to the base cytosine when it is immediately followed by guanine — by a process known as DNA methylation (see graphic). The methyl group can be sensed by proteins that turn gene expression on or off through regulating chromatin structure. The second, more complex kind of alteration involves changes to the histones around which chromosomal DNA is wrapped. Each histone has a protruding 'tail' to which more than 20 chemical tags can attach, like charms on a bracelet. Some of these tags, or certain combinations of them, dubbed the histone code, give rise to relaxed chromatin; others have the opposite effect.

Epigenetic codes are much more subject to environmental influences than the DNA sequence. "This could explain how lifestyle and toxic chemicals affect susceptibility to diseases," says Vardman Rakyen, a researcher at the Sanger Institute. "Up to 70% of the contribution to a particular disease can be non-genetic." Indeed, one key finding of Esteller and his team's study was that epigenetic profiles of twins who had been raised apart or had noticeably distinct lifestyles differed more than



The two main components of the epigenetic code

DNA methylation

Methyl marks added to certain DNA bases repress gene activity.

Histone modification

A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

those who had lived together for a while or shared similar environments and experiences¹. Rakyen himself is studying a cohort of identical twins, where one twin has type 1 diabetes and the other does not, with the aim of finding epigenetic changes associated with the disease.

Although different labs around the world, such as Rakyen's, are already working on their own individual studies, several researchers argue that it is time for a coordinated effort. A series of international workshops, and expert and government reports have emerged in recent months that address the value and scope of an international human epigenome project. The ultimate goal of such a project would be to identify all the chemical modifications of DNA and histone proteins for all chromosomes in all types of normal human tissue.

As for the HGP, an international consortium

would set priorities, coordinate research efforts, centralize materials and resources, create the necessary technologies and monitor research progress.

Piece by piece

"Epigenomics is at a stage where genomics was 30 years ago, when everyone was working on their part of the puzzle," remarks cancer biologist Peter Jones at the University of Southern California, Los Angeles. Jones was formerly president of the American Association for Cancer Research (AACR), which is based in Philadelphia. "We need to see the bigger picture. It takes concerted efforts on an international scale. And this is how the IHEP would make a difference."

Although a number of funding bodies — such as the Wellcome Trust, the AACR, the US

TAGGED FOR DISEASE

Thanks to its ability to instruct cells how to 'play' the genetic instructions spelled out in their DNA, the epigenetic code is proving to be central to processes such as development, ageing, cancer, mental health and infertility. Cancer researchers, in particular, are studying it to develop diagnostic and prognostic tools and drugs.

Epigenetics may contribute as much to cancer development as mutations in the DNA itself. For example, *Wnk2*, a gene whose activity is thought to suppress tumour development, is more often shut down by epigenetic

changes in certain brain tumours than it is lost by genetic deletions.

Epigenetic marks can also help predict clinical outcomes. Researchers have identified a pattern of epigenetic modifications within the oestrogen receptor gene that correlates with a cancer patient's chances of survival in response to treatment with the drug tamoxifen⁹.

Epigenetic changes are easier to reverse than genetic mutations, by adding or removing the chemical tags involved. Drugs that do this are now at the forefront of cancer treatments. For example,

Vidaza, a drug made by Pharmion Corporation of Boulder, Colorado, has been approved by the US Food and Drug Administration for the treatment of myelodysplastic syndromes, also known as 'preleukaemia'.

In theory, this drug, which blocks DNA methylation, should be active in all cells. In practice, it seems to be active only in cancerous cells, stimulating the expression of many genes, including those that suppress cancer development. A number of other epigenetic drugs are at various stages of clinical trials.

J.Q.

the lesser of two evils, but exactly how different the epigenomes of cell lines are compared with normal tissues remains to be seen. The inclusion of cell lines in some pilot studies in the proposed IHEP should be able to resolve this issue.

Final frontier

Perhaps the greatest challenges facing the IHEP are technological: mass-production-style tools must be developed to decode the epigenome, and the morass of data will have to be stored and analysed. At the moment, the main method used to determine DNA methylation sites is reliable, but extremely expensive, and the technology used to study histone marks is prone to problems with accuracy and reproducibility.

Scientists hope to tackle these problems by linking the IHEP to projects on the epigenomes of lab workhorses, such as the yeast, fruitfly and mouse, for which techniques are more advanced. Computational scientists are also developing the sophisticated bioinformatics tools needed to store and analyse multi-dimensional epigenome data.

Given these technological challenges, it is only natural to question whether the research community is ready for such an enormous undertaking. Drawing on the experience of the early days of planning for the HGP, researchers working on epigenetics are unanimous in thinking they can do it. They have drawn up a plan of how to manage the international assets available for the

IHEP^{5,7-9} and say that, like the HGP, the IHEP will catalyse its own development. "One can never be 100% ready. We have 60% of the technology to go for the real thing," says Thomas Jenuwein, a molecular biologist at the Research Institute of Molecular Pathology at the Vienna Biocentre, Austria. "The rest will happen once the momentum is built up. We should have that vision to go in big."

Jane Qiu is an editor at *Nature Reviews Neuroscience*.

1. Fraga, M. F. *et al. Proc. Natl Acad. Sci. USA* **102**, 10604-10609 (2005).
2. Waddington, C. *Endeavour* **1**, 18-20 (1942).
3. Waterland, R. A. & Jirtle, R. L. *Mol. Cell. Biol.* **23**, 5293-5300 (2003).
4. Weaver, I. C. G. *et al. J. Neurosci.* **25**, 11045-11054 (2005).
5. Lieb, J. D. *et al. Cytogenet. Genome Res.* **114**, doi:10.1159/000091922 (2006).
6. Raky, V. K. *et al. PLoS Biol.* **2**, e405 (2004).
7. Feinberg, A. P. & Jones, P. A. Workshop on Defining the Epigenome <http://dcb.nci.nih.gov/Workshoprpt.cfm> (2006).
8. Rauscher, F. J. III *Cancer Res.* **65**, 11229 (2005).
9. Jones, P. A. & Martienssen, R. *Cancer Res.* **65**, 11241-11246 (2005).

National Cancer Institute (NCI), and the US National Human Genome Research Institute (NHGRI) — have shown interest by taking part in the discussion, funding agencies have yet to commit to financing and leading the project.

All aboard

Since the completion of the Human Genome Project, there have been many multi-centre schemes, each of which costs millions or even billions of dollars. Some of these initiatives, such as the US National Institutes of Health Human Cancer Genome Atlas (which aims to identify and catalogue genetic mutations in human cancers), have prompted arguments over scale and cost-effectiveness. A key question for funding bodies is whether the IHEP would be yet another multi-million-dollar project. Proponents say no. "The goal of the IHEP is not to create another big enterprise, but to make things as cost effective as possible, to interface with wonderful projects that are under way and to fund important pilot projects," says Andrew Feinberg, director of the Centre for Epigenetics of Common Human Disease at the Johns Hopkins University in Baltimore, Maryland.

A number of smaller scale multi-centre epigenome projects are already under way or under discussion in Europe, the United States, India and Japan. Most prominent is that set up by the European Human Epigenome Project (HEP) Consortium in 2000. Following the publication of a pilot project in 2004 (ref. 6), the European HEP Consortium will soon make its data on the epigenetics of the entire

chromosomes 6, 20 and 22 publicly available.

Although few people doubt the importance of an international human epigenome project, how to go about it remains a subject of debate.

A key challenge is defining what the epigenome entails and what cell types to study. Some researchers argue that the project should first tackle blood cells, because they are easy to collect and work with, and are our main 'window' into the epigenome of both healthy and diseased individuals. Once a high-resolution blood epigenome is determined, it will serve as a reference with which other epigenomes, including those of diseased or ageing tissues, could be compared.

But the diversity of epigenomes in different cell types means that it may not make sense to restrict pilot projects to one single tissue, or to a particular time in a tissue's development. After intense discussion in three recent international workshops^{5,7-9}, researchers in the epigenetic community now agree that initially eight to ten tissues, including the blood, should be studied simultaneously. Ultimately, the epigenome of all tissues, including embryonic stem cells, will be mapped out.

Another question is whether to study cells grown in the lab or biopsies of tissues taken from people. Biopsies contain different cell types, which would muddy the picture, but lab-grown cells might contain abnormal epigenetic tags. At the moment, some biologists are leaning towards lab-grown cells as being

"Epigenomics is where genomics was 30 years ago, when everyone was working on part of the puzzle."

— Peter Jones

The mountain Lappporten looms over the village and research station of Abisko, Sweden.

ON THIN ICE

The Arctic is the bellwether of climate change, which shows up there first and fastest. **Quirin Schiermeier** visits ecologists struggling to keep up.

On a splendid, freezing spring morning in Sweden's far north, global warming seems far away. The pristine landscape around the Abisko Scientific Research Station, 200 kilometres inside the Arctic Circle, is glistening white. Thick ice covers nearby Lake Torneträsk.

The spring equinox has passed, and the days are quickly getting longer. But it is -15°C as Gareth Phoenix, a plant ecologist from the University of Sheffield, UK, who has wintered at the station, wades outside to check a 'snow melt' experiment installed between mountain birches.

Thin heating cables and four 1,500-watt lamps hanging half a metre above the ground have melted the snow in small patches, exposing shrubs and lichens. Phoenix looks pleased. "Warming the Arctic outdoors in winter isn't terribly easy," he notes. With Massachusetts-based engineer Frank Bowles, Phoenix has spent weeks tinkering with the array so that it melts snow without toasting the plants. He thinks he's got it right now: the heat has melted 45 centimetres of snow in three-and-a-half days, exactly what can happen during extreme warming events in the Arctic winter.

Such warming events occur much more frequently at Abisko now than at any time since climate records began there in 1913, says the centre's scientific director, Terry Callaghan. Ecologists are worried that the short-lived melting episodes, and the sudden return to cold weather afterwards, could harm plants and soils. Such disturbances could resonate through the whole ecosystem in such areas, they suspect, with potentially devastating knock-on effects for nutrient supply, plant growth and animal populations.



Change comes faster in the Arctic than elsewhere. As the snow-free season lengthens and sea ice becomes less abundant, albedo — the proportion of sunlight reflected by the ocean and the ground — decreases. The sunlight that's not being reflected by snow and ice is absorbed instead, and this amplifies climate warming at high northern latitudes. Computer models and on-the-ground observations both suggest that the most pronounced warming will occur in winter¹.

At the Abisko station, mean temperatures between December and February have risen by around 5.5°C over the past century — eight times more than the average rise in the Northern Hemisphere². Similar warming has been observed throughout the Arctic, causing glaciers in Greenland to flow faster, permafrost soils in Siberia to thaw, and boreal forests in

Russia and Canada to move further north.

The Abisko station has become a prime location for studying the effect of climate change on terrestrial Arctic ecosystems. Streams of very warm air masses have been observed here at least once every winter over the past seven years, causing temperatures near the surface to rise by 25°C or more within a day. The warm air melts the snow cover, exposing and sometimes killing the plants beneath. If the melted snow then refreezes, the shrubs and lichens become encrusted with ice and are no longer accessible as food for lemmings and reindeer. Recent population crashes of wild reindeer on the Arctic island of Svalbard are thought to be linked with such 'icing' events, although a connection is not proven.

Unknown unknowns

Experiments such as the snow-melt study, which will run for at least three years, are meant to clarify the short- and long-term effects of the melting episodes. Mimicking nature is not nature itself, however: Phoenix says that much more study is needed before scientists can hope to understand the complexity of the changes going on. "The problem is to find out what the most interesting thing is of what you are measuring," he says. "You often don't know what you don't know."

Each year, around 700 scientists and students come to Abisko to study Arctic climate and environment, carbon cycles, lake ecosystems and geomorphology. The station has plenty of rooms and lab space, as well as a new scenic lecture theatre and conference facility. And unlike remote Arctic research bases, Abisko is wired with electrical power. This

luxury not only makes a stay at the station a comfortable experience; it is also a prerequisite for running long-term outdoor experiments that require a constant energy supply. "It would hardly be possible to run a winter snow-melt experiment at a place like Toolik Lake station in Alaska, where you have only diesel generators," says Jerry Melillo, co-director of the Ecosystems Centre at the Marine Biological Laboratory in Woods Hole, Massachusetts. For his part, Callaghan argues that many more long-term stations, like Abisko, are needed for the monitoring efforts (see pages 127 and 133).

From sink to source

A key problem, he says, will be to determine whether the Arctic, which currently accounts for one-third of soil carbon storage on Earth, is likely to remain a carbon sink, or whether it will turn into a source of carbon. As soils grow warmer, many worry that greater microbial activity could increase the rate of decomposition and lead to increased releases of methane and carbon dioxide³.

For the moment, computer models suggest that the Arctic is still a small carbon sink. But the trends are highly inconsistent, says Callaghan. Most Arctic lakes, for example, seem to be saturated with carbon dioxide and have turned into carbon sources.

At Abisko, tall measurement towers and small chambers around individual plants monitor the carbon flowing from soils and vegetation to the atmosphere and back again. Such data are then fed into complex carbon balance and vegetation models. But adding carbon-flux data from just one additional site can have a huge impact on the overall picture.

At Abisko, Callaghan has seen how small perturbations can affect the carbon balance of a whole region. From his office window, he points at a bare slope on the distant shore of Lake Torneträsk. During the exceptionally warm winters of 1950 and 2004, eggs of the autumn moth (*Epirrita autumnata*), a caterpillar feeding on mountain birch, survived there in vast numbers. Later in the year the insects destroyed large swaths of the forest,



Thaw point: an experiment to study what happens to plants when snow cover melts and refreezes.

which has not recovered since. Insect outbreaks such as this can convert a whole area from a carbon sink into a source.

Climate warming sometimes brings what may look like good news. In Lapland, for example, the tree line has risen 60 metres since 1900, and satellite images confirm that forests and shrublands have also increased in the northern parts of Siberia, Canada and Alaska. The extra biomass, some believe, could suck up more carbon dioxide.

But once again, the picture isn't simple. Because trees decrease albedo in comparison with the tundra vegetation they replace, their spread might actually accelerate warming in Arctic regions. A recent study in Alaska suggested that terrestrial changes in summer albedo plus lengthening of the snow-free season already has an effect similar in magnitude to the warming expected from a doubling of carbon dioxide. And as shrubs and trees continue to proliferate, as some models predict they will in a carbon dioxide-rich world, they could further amplify Arctic warming by two to seven times⁴. "Despite

what governments may say, forest growth may do more harm than good to the climate," says Callaghan.

The Abisko station has taken a lead role in the European Union-funded BALANCE project, which investigates present and future climate-change vulnerabilities in the region.

In so doing, the researchers at Abisko have turned to some valuable allies: the local Sami population of reindeer herders, whose indigenous knowledge of the area could help scientists assess changes in terrestrial

Arctic ecosystems.

"The Sami are all around the landscape and they see many things we don't," says Callaghan. "They are really the missing link between our individual observations and satellite imagery. And including indigenous knowledge just makes our own experiments more relevant." For instance, this spring saw the launch of a joint research project, including linguists, anthropologists and Sami academics at the Nordic Sami Institute in Kautokeino, Norway. 'Snow and Ice' will assess the impact of environmental changes and extreme events on reindeer herding and the movement of reindeer and people through the region.

Lessons from the winter snow-melt experiment are just one thing scientists at Abisko hope to share with the Sami. As night falls over Lake Torneträsk, the eerie green veils of the aurora borealis drift across the starry sky. Relaxing in the improvised 'tundra bar' in the station's cellar, Callaghan is scheming about supplying the Sami with small, portable weather stations. Then they too could become part of Abisko's science network. ■
Quirin Schiermeier is *Nature's* German correspondent.



Herding reindeer has given Norway's Sami population an intimate knowledge of their environment.

1. Arctic Climate Impact Assessment *Impacts of a Warming Arctic* (Cambridge Univ. Press, 2004).
2. Callaghan, T. et al. *Polar Research* (in the press).
3. Mack, M. et al. *Nature* **431**, 440–443 (2004).
4. Chapin, F. et al. *Science* **310**, 657–660 (2005).

BUSINESS

Challengers in the field

The agricultural biotechnology industry is ten years old, and the story of its first decade has largely been about one company selling crops with single transgenic traits. Monsanto, based in St Louis, Missouri, still dominates the sector — but that could be about to change.

A deal struck last month between Monsanto's two fiercest rivals will, they hope, presage a second era characterized by a more even spread of market share and the production of multiple traits 'stacked' into single plants.

Syngenta, based in Basel, Switzerland, and Pioneer Hi-Bred of Des Moines, Iowa — a subsidiary of DuPont — last month formed a 50–50 venture which, they believe, will give Monsanto a run for its money.

The venture, GreenLeaf Genetics, is based in Omaha, Nebraska, and plans to license traits from both Syngenta and Pioneer to the host of small, long-established local companies that supply many farmers with corn (maize) and soya-bean seeds.

"For the farmer, this means he'll be able to get the benefit of technology from both of these companies from his local seed company," says Ron Wulfkühle, GreenLeaf's president.

The approach is a change of direction for Pioneer, the world's oldest and largest seed company, which until now has sought to sell its transgenic seed directly to farmers. Plant strains with stacked traits will follow as new traits become available over the next few years. "GreenLeaf will allow us to present a broader suite of products," says Pioneer's president, Dean Oestreich.

About one-third of America's \$2.8-billion corn-seed market is supplied through independent companies, and Monsanto has scored major success by licensing its transgenic technology to them. It also sells direct and licenses genes to rivals, including Pioneer — with the result that up to four-fifths of all transgenic corn and soya beans contain Monsanto traits.

GreenLeaf's challenge is to break this



Growth industry: transgenic crops with multiple traits are catching on in North America.

dominance. Its initial emphasis will be on corn for North America, which accounts for more than half of the total area of genetically modified (GM) crops planted worldwide, according to the ISAAA, a Philippines-based organization that promotes the technology in developing countries.

GM seed now produces about 60% of the world's soya beans and 16% of its corn. The technology has yet to make substantial inroads in wheat or rice and, although genetic modification has spread rapidly (see graph), consumer resistance has held it back in places. There is no cultivation of GM crops in Japan or Britain, very little in the rest of Europe, and the only transgenic crop grown in China is cotton.

Clive James, a prominent plant scientist and chair of the ISAAA, predicts that the total planted area of GM crops will grow from 90 million hectares this year to at least 200 million hectares in ten years' time. If transgenic rice takes off in Asia, he suggests, that alone could add 250 million hectares or more.

But perhaps the most significant development is the increasing spread of multi-trait crops. Most of these seeds currently have two traits, although future plants could incorporate resistance to weedkillers and to drought with insecticides, for example, and with specific nutritional benefits.

GreenLeaf will make multiple traits available to seed companies, says Wulfkühle. One such trait will be higher-yield herbicide resistance based on Pioneer's Optimum GAT technology, which makes plants resistant to the weedkiller glyphosate and will be ready in 2009.

James also expects to see the commercial arrival of traits that are more useful to farmers in poor countries. That would answer critics who say early traits only addressed the demands of the world's richer farmers.

Colin Macilwain

IN BRIEF

COPYCAT RIGHTS Two firms that make generic drugs have received a windfall courtesy of a US district court. Teva Pharmaceutical Industries in Israel and Ranbaxy Laboratories in India were granted exclusive sales rights for six months to generic versions of Zocor (simvastatin). Made by Merck, Zocor is the world's second-largest cholesterol drug in terms of sales — and it comes off patent next month. The court in the District of Columbia ruled that the Food and Drug Administration (FDA) acted unlawfully when it denied the two firms six months of exclusivity on the drug. The FDA may appeal.

INDUSTRIAL SLOWDOWN

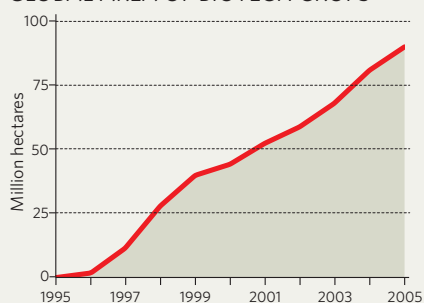
Industrial support for research at US universities dropped for the third successive year in 2004, falling by almost 3% to \$2.1 billion, according to the National Science Foundation. For every \$20 spent by academics in science and engineering that year, industry provided \$1. The drop is a result of industry becoming "much more short-term and development-oriented" after the US economy dipped in 2001, says Kei Koizumi, an analyst at the American Association for the Advancement of Science.

LILLY THINKS SMALL A division of drugmaker Eli Lilly has signed a collaborative agreement with Altair Nanotechnologies, a Nevada-based company that supplies ceramic nanomaterials and technologies. In exchange for undisclosed milestone and royalty payments, Indiana-based Elanco Animal Health will get exclusive rights to use Altair products to develop and deliver animal drugs. Altair's materials have already been used to develop RenaZorb, a drug for kidney dialysis patients, which it licensed to Spectrum Pharmaceuticals in January.

RIGHT TO TRY A US patients' rights group is making progress in its fight to win dying patients the right to try drugs still under development. The Abigail Alliance, which lobbies for patients with terminal diseases, sued the Food and Drug Administration in 2003, claiming that the terminally ill have the right to use experimental drugs that have passed initial human-safety trials. A lower court threw out the case in 2004, but a higher court last week reinstated it.

SOURCE: ISAAA

GLOBAL AREA OF BIOTECH CROPS



Medical council funds both clinical and basic research

SIR — Your Editorial “Brown’s budget briefing” (*Nature* **440**, 581; 2006) implies that the UK Medical Research Council (MRC) is responsible only for basic biomedical research. In fact, MRC scientists invented randomized clinical trial methodology in the 1940s, and the MRC remains the largest UK public funder of trials. It has supported a multitude of trials including the Heart Protection Study (demonstrating the value of statins) and the CRASH trial (showing that the treatment of head-injury victims with corticosteroids is actually dangerous), both of which were highlighted in the budget statement.

The MRC is internationally recognized for supporting research that has underpinned the twentieth-century revolution in basic biomedicine. But its contribution to clinical research has been equally significant. Not just its trials and epidemiology (including discovery of the link between smoking and cancer), but ground-breaking work in such areas as vitamins, viruses, penicillin, vaccines, magnetic resonance imaging and antibody drugs.

Three years ago, the MRC made a strategic commitment to increase its investment in translational, clinical and public-health research. The planned creation of a new, single fund for health research is an opportunity to strengthen further the pipeline from discovery to better treatment and prevention. But, as you point out, maintaining the independence of scientific judgement that has delivered so well in the past will be vital.

Colin Blakemore

Medical Research Council,
20 Park Crescent, London W1B 1AL, UK

Exaggerating one's success is rife in Chinese academia

SIR — Your News in Brief story “Fantasy reference list leads to the sack” (*Nature* **440**, 728; 2006) reports on the dismissal of Hui Liu, assistant dean of the medical school at Tsinghua University in Beijing. But exaggerating the scientific significance of one’s research is even more harmful to the reputation of China’s academic communities than faking a CV or résumé.

The Chinese Academy of Sciences (CAS) plays an important role in setting the course for scientific and technological development. Its members are scientists, engineers and technologists, who are eminent scholars with a strong influence on China’s science and technology policy. CAS membership is a high honour, but reading the exaggerated

biographies of some members and candidates, one feels some of them deserve Nobel prizes. Ordinary research results are often portrayed as extraordinary scientific achievements.

Partly being driven by the country’s ‘knowledge innovation’ policy, Chinese scientists have made it a priority to pursue publications in journals with high impact factors. The peer pressure on postgraduate students, as well as on CAS members, has made exaggeration of scientific achievements and overstatement of personal talent quite common in China. But ill-minded over-exaggeration, bragging, plagiarizing and fraud are morally wrong and should be legally challenged, in China or anywhere else.

Zheng Huang

Radiation Oncology Department, University of Colorado at Denver and Health Sciences Center, Aurora, Colorado 80045, USA

Allergy test might have avoided drug-trial disaster

SIR — Your News story “Can super-antibody drugs be tamed?” (*Nature* **440**, 855–856; 2006), on the TGN1412 anti-CD28 antibody drug and the fallout from this near-fatal clinical trial, lacks a simple interpretation of the data based on common sense.

Accounts of the clinical trial describe how, following intravenous infusion of the antibody, the volunteers were clawing off their clothes, swelling with obvious oedema, and starting to suffer pain and panic within minutes. The developer of TGN1412, Thomas Hunig, links this disabling “cytokine storm” to T-cell activation. But such a reaction would take hours.

In fact, these volunteers could have been victims of the release of preformed mediators, of the types usually found within the granules of cells of an allergic response. As any subject of an anaphylactic response can testify — myself included — the response is immediate, with similar symptoms to those experienced by the volunteers in the TGN1412 trial.

The literature suggests that CD28 can be expressed by cells of the allergic response (M. Tashiro *et al.* *J. Immunol.* **158**, 2382–2389; 1997), and the real disaster was that no tests were carried out for the possibility of allergic reactions before TGN1412 was released into the blood stream. A simple ‘weal and flare’ allergic-response test could have easily alerted the clinicians to a potential allergic reaction. But reports have suggested that no such tests were performed. The first human trial with this drug was the intravenous infusion that led to such catastrophic responses in these volunteers.

John H. Weis

Pathology Department, University of Utah School of Medicine, 30 North 1900 East 5C124 SOM, Salt Lake City, Utah 84132, USA

Keeping an eye on privacy issues with geospatial data

SIR — In their Commentary “Mapping disaster zones” (*Nature* **439**, 787–788; 2006), Illah Nourbakhsh and colleagues rightly say that the global distribution of spatial data, especially high-resolution aerial photography, may raise issues of privacy violation. We at Louisiana State University used Google Earth extensively in the Emergency Operation Center to facilitate search-and-rescue missions immediately after Hurricane Katrina. There is no denying the utility this system provided. But, in the months after the hurricane, we used the markings left by these same search-and-rescue teams to highlight the danger of using geospatial technology.

A map displaying body-recovery locations in the New Orleans area was cut out of a local newspaper and turned into a Geographic Information System (GIS) layer, using census boundary information to guide the transfer as no streets and limited city landmarks were on the original. Once in the GIS, the central coordinates of each mortality ‘dot’ were extracted. A field team using the Global Positioning System checked the locations where bodies had been found against the search-and-rescue markings sprayed on houses. In many cases the dots, which on the map covered approximately one-and-a-half city blocks, revealed the location to be within an area of two or three neighbouring houses, with the exact house being identified in some cases.

What would have happened if a scientist had presented HIV cases, or nesting sites of an endangered species, on a similar city display — relying on an oversized point symbol to mask the real-world location? There is scant guidance for scientists to help preserve spatial confidentiality. Current standards often address only the required size for denominator populations in a thematic map display. Unless we in academia take the lead in expanding and enforcing a more rigid set of spatial display rules, especially for point data, we run the risk of an over-zealous tightening of data release and a protracted battle to again persuade those in power that a map can be used for the good of society.

Andrew Curtis*, **Jacqueline W. Mills†**,
Michael Leitner*

*Department of Geography and Anthropology, Louisiana State University, Baton Rouge, Louisiana 70803, USA

†LSU GIS Clearinghouse, CADGIS Research Lab, Louisiana State University, Baton Rouge, Louisiana 70803, USA

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

SPRING BOOKS

Evolution of the selfish gene

The thirtieth anniversary of Richard Dawkins' landmark work provides an opportunity to take stock.

Richard Dawkins: *How a Scientist Changed the Way We Think*

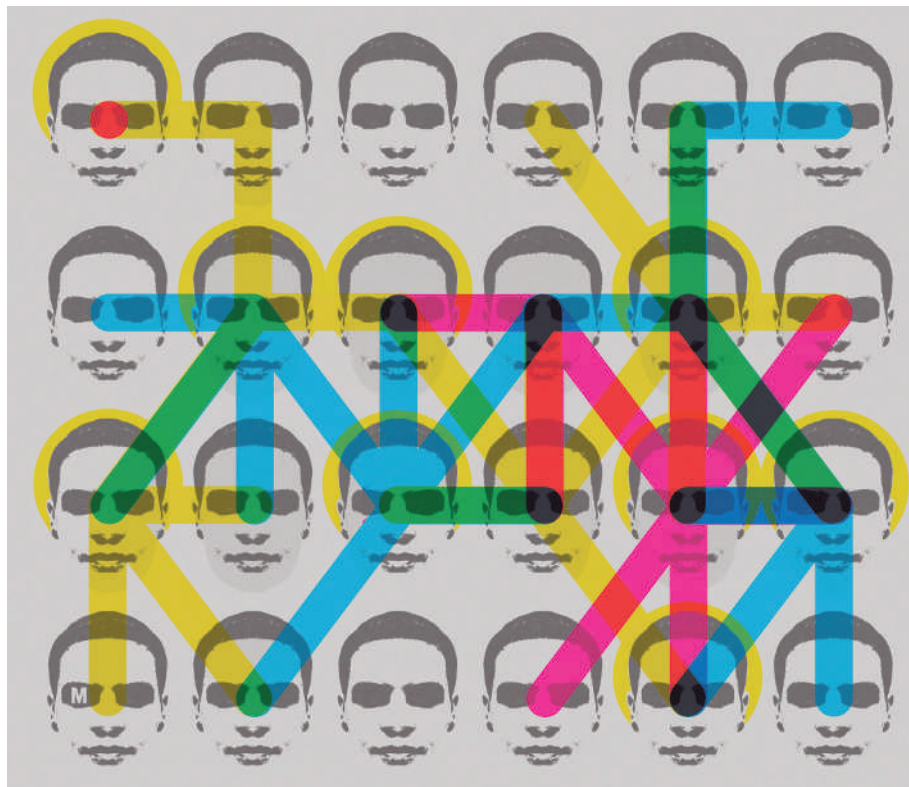
edited by Alan Grafen & Mark Ridley
Oxford University Press: 2006. 304 pp.
£12.99, \$25

Dan Sperber

"We are survival machines — robot vehicles blindly programmed to preserve the selfish molecules known as genes. This is a truth which still fills me with astonishment ... One of my hopes is that I may have some success in astonishing others." That hope, expressed by Richard Dawkins in the preface of *The Selfish Gene*, has been more than fulfilled. Published 30 years ago, *The Selfish Gene* has been, and remains, one of the most influential science books of all time. To celebrate this anniversary, a third edition has been released, along with *Richard Dawkins: How a Scientist Changed the Way We Think*. The latter is a collection of comments and testimonials edited by Alan Grafen and Mark Ridley.

In the 1960s and '70s, biologists such as William Hamilton, John Maynard Smith, Robert Trivers and G. C. Williams sought to explain troubling aspects of evolution by approaching them at the level of genes, rather than at the usual level of individual organisms, groups or species. The existence of altruistic behaviour that may decrease the reproductive success of altruistic individuals, for instance, presents a puzzle for standard darwinism. Hamilton showed that altruistic behaviour may increase the chances of the genes involved being replicated when the beneficiaries are also carriers of the same genes; this gene-level selection explains the evolution of altruism.

What Dawkins did was integrate such findings into a vivid and systematic picture of biological evolution wholly "from the point of view of the gene" and explore the wider implications of this approach. His picture challenges common-sense ontology and expectations, and is indeed astonishing. We spontaneously interpret the behaviour of individuals as that of agents capable of pursuing their interest, and extend this kind of interpretation to social groups. Fragments of molecules, on the other hand, are unfamiliar objects to which we are not disposed to attribute interests and goals. Of course, genes are not literally agents, let alone selfish ones intent on propagating themselves, but analysing what they would do if



ILLUSTRATIONS BY JOE MAGEE

they were provides us with a uniquely cogent account of their actual effects on the world. The logic is that used by Darwin when he explained the existence in nature of design without a designer as an effect of selection. Dawkins shows how this logic can be exploited at the micro-level of 'replicators', or genes.

Whether evolution and selection are best described at the level of genes, or at the higher levels of organisms or groups, or at all these levels simultaneously, remains contentious (with echoes of the debate in Grafen and Ridley's edited volume, *Richard Dawkins*). There is little doubt, however, that the gene-centred approach has been the source of novel and deep insights. In particular, further challenging common sense, Dawkins attacked the view of organisms as bearers of life *par excellence*. Only in some cases do different replicators cooperate in such a way that larger coherent units — the 'vehicles' of replicators — emerge. Not all these vehicles correspond to the individual organism. In *The Extended Phenotype*, his second book and possibly his best, Dawkins showed how the vehicles may extend

beyond the boundaries and behaviour of organisms, for example when the genes of parasites express themselves by modifying their host's behaviour.

In the last chapter of the first edition of *The Selfish Gene*, Dawkins introduced what may be his most popular idea, that of 'memes', or cultural replicators. Any population of entities that produce copies of themselves, and that vary both in their specific features and in their reproductive success, are replicators and are thus candidates for darwinian selection. The idea that cultural evolution might be modelled along darwinian lines had often been suggested, but Dawkins reformulated it with characteristic crispness and clarity. Contrary to the idea that cultural items (such as ideas, skills, practices and artefacts) thrive because of their contribution to the social or biological welfare of the individual or groups that adopt them, Dawkins argued that cultural items may thrive because they cause their own propagation. Establishing the possibility of such 'selfish memes' served two purposes: generalizing the idea of replicators

beyond biology, and suggesting an evolutionary approach to culture.

However, as Robert Aunger observes in Grafen and Ridley's book: "No significant body of empirical research has grown up around the meme concept ... In fact the memetic literature remains devoted almost exclusively to theoretical antagonisms, internecine battles, and scholastic elucidations of prior writing on memes." Dawkins has essentially left the development of memetics to others. Instead, together with Daniel Dennett, he has used the idea of the meme as a powerful tool in his criticism of religious ideas, which he describes as "viruses of the mind". The effectiveness of the criticism does not much depend on the scientific details of a would-be memetics.

Dawkins' depth and clarity of vision, intellectual honesty and passion, and superb writing have indeed changed the way many of

us think about biology. And the latest volume, *Richard Dawkins*, brings together testimonials and reflections about Dawkins himself or inspired by his work. Most of the contributions, by eminent scientists, philosophers and writers, are laudatory; a few are critical. The book is a pleasant read and throws useful light on the multifarious impact of Dawkins' work on biology, philosophy, science writing and the public debate on science and religion. Particularly illuminating are Grafen's chapter discussing the relationship between Dawkins' work and more mathematically oriented population genetics, and Ullica Segerstråle's chapter on Dawkins and sociobiology. Still, in preparing this review, I re-read *The Selfish Gene*, and this was the real treat. ■

Dan Sperber is at the Centre National de la Recherche Scientifique and the Institut Jean Nicod, 1 bis avenue Lowendal, 75014 Paris, France.

novels. Yet, perhaps because of the intricate mixture of paracelsian magic, metallurgy, medicine and alchemy, the historical Paracelsus has received comparatively little attention. By presenting the work of Paracelsus, including all the contradictions and neologisms, as an intensely personal enterprise embedded in Renaissance life, Ball circumvents many of the historical difficulties and comes up with an excellent biography that is relevant for historians and general readers alike.

Ball takes events in the life of Paracelsus as starting points for discussing the Renaissance world. For instance, when discussing Paracelsus' life as a vagabond, Ball speaks about the difficulties of travelling in early modern Europe. His discussion of the alchemy of Paracelsus transforms into a discussion of economic growth and the power of miners, and his religious and political views are compared to those of reformers and princes. Ball speaks of Paracelsus' views on astrology in relation to the astronomy of Copernicus and his followers. And last but not least, Ball writes extensively about the traditional (galenic) medicine and chemistry that Paracelsus challenged. The book's illustrations provide a vivid picture of the time and further enliven Ball's account.

This approach is brave and enriching but is also a little overwhelming. At times Paracelsus disappears into the background, and the reader is in danger of getting lost in detailed descriptions of Renaissance culture. Moreover, in a book as ambitious as this it is almost unavoidable that the terminology becomes at times confusing. Much of Paracelsus' work teeters on the brink of the spiritual, and his own vocabulary often seems puzzling to modern readers.

The difficulty in understanding Paracelsus' neologisms and expressions is clearly visible in this book. Ball, for instance, is often wobbly in calling details of Paracelsus' work 'mechanical', 'spiritual' or 'materialistic'. Sometimes it is not clear whether he adopts Paracelsus' own words or gives them a modern, and therefore different, meaning. For example, Ball maintains that Paracelsus' concentration of nature's potencies in the preparation of medicines was "not mechanical" (presumably in a modern sense), but on the next page he states that, according to Paracelsus, the powers of the stars permeate "mechanically" through the Universe (thereby referring to Paracelsus' own words). In both cases Ball refers to the working of invisible powers, but apparently these are mechanical in one case but not in the other, leaving it unclear what Paracelsus meant when he spoke of the mechanical working of the invisible forces of nature. The same goes for the important paracelsian distinction between the material and the spiritual, which at times makes Ball's description of Paracelsus' thoughts somewhat bewildering.

To be fair to Ball, he does explain many of Paracelsus' neologisms, but he also has the

Renaissance magic and mysticism

The Devil's Doctor: Paracelsus and the World of Renaissance Magic and Science
by Philip Ball

Heinemann/Farrar, Straus, Giroux: 2006.
448 pp. £20/\$26

Rina Knoeff

One has to admire Philip Ball's courageous undertaking in writing a biography of Paracelsus, arguably the most controversial medical writer in the Renaissance. Not only are the works of Paracelsus' own hand extremely difficult to read and understand but, more importantly, historical reconstructions of his life and thoughts complicate the picture to such an extent that it is hard to write a 'fair' biography.

Paracelsus is known for being a failed physician; a psychiatric subject in the casebooks of the psychoanalyst Carl Jung; a German national hero during the Nazi period; and the founder of biochemistry. Ball, however, sets Paracelsus in the social, religious and cultural life of his time, a refreshing move away from the tendency to describe early 'scientists' as the forerunners of today's scientific developments.

Ball is aware of the historiographic difficulties surrounding the life and work of Paracelsus. His account starts with a brief discussion of how magic and

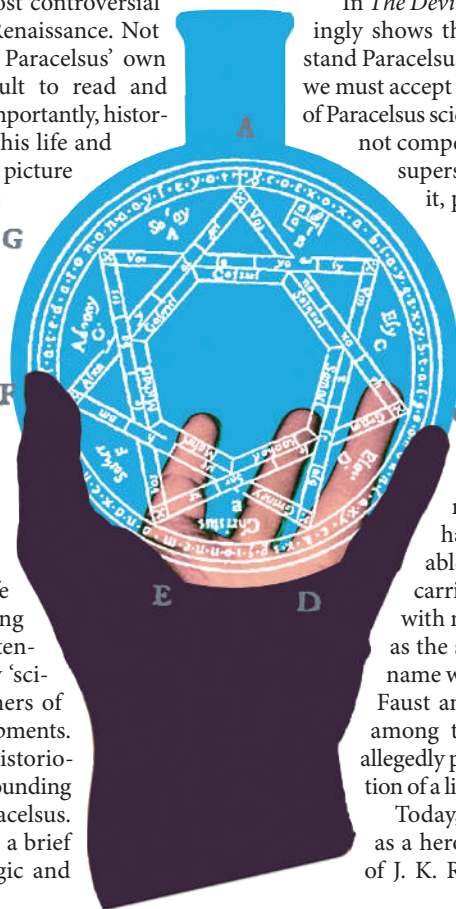
occult forces were accepted parts of early modern science. In addition, Ball acknowledges the close connection between early modern natural philosophy, Renaissance humanism and Reformation religion. In so doing, he follows a fairly recent trend in the history of science and medicine in which religion and science are seen as mutually shaping fields of knowledge.

In *The Devil's Doctor*, Ball convincingly shows that in order to understand Paracelsus' work and personality we must accept that "in the philosophy of Paracelsus science and rationalism do not compete with mysticism and superstition but blend with it, producing a world that

now seems at the same time wonderful and bizarre".

Paracelsus, or Philippus Aureolus Theophrastus Bombastus von Hohenheim, spoke to the imagination. He is said to have ridden a magical white horse, to have cured many incurable diseases, and to have carried an enormous sword with magical powers, as well as the secret elixir of life. His name was linked with those of Faust and Martin Luther, and among the many miracles he allegedly performed was the creation of a living, human-like being.

Today, Paracelsus appears as a hero in the magical world of J. K. Rowling's *Harry Potter*



tendency to follow Jung in maintaining that the language used by Paracelsus must be seen symbolically as an expression of his unconscious mind. It is too easy to argue that, for this reason, the lexicons provided by scholars of Paracelsus must be taken “with a pinch of Paracelsian salt”. To my mind, one of the most urgent tasks of the historian is to find out what precisely Paracelsus meant. Without making this effort, the description of his work can only remain superficial.

This criticism notwithstanding, *The Devil's Doctor* is a fascinating read, rich in content and hugely entertaining. Moreover, it shows that magic was as much at the root of modern science as were the famous discoveries of our modern scientific heroes. It is this awareness that makes Ball's account of Paracelsus essential reading for historians and scientists alike.

Rina Knoeff is in the Faculty of Arts, University of Leiden, 2300 RA Leiden, the Netherlands.

sent that afternoon to Rubin and Francis Collins have passed into legend. Rubin was found, Venter confronted. To the fury of Celera's business-development team, Venter backed down, and within 24 hours the fly sequence was freely available on GenBank. Arguably, this was the moment when it became clear that nobody would make serious money from genome sequencing; the pressure to share would prove irresistible. Indeed, given the way the world was going that autumn, maybe it was an omen for the whole 'dot com' frenzy in whose slipstream Celera's share price had accelerated. Celera was not the only overvalued company with a flawed business model for making money out of digital information.

Not that Ashburner portrays himself as a hero or Venter as a villain. The delightful sequel is that within a week of this stressful incident, Ashburner was at the Celera jamboree playing happily in the toybox that was the fruitfly genome, exhilarating in the new knowledge and even joshing with Venter. At one point, while celebrating the end of the

The soldier's tale

Won for All: How the *Drosophila* Genome was Sequenced

by Michael Ashburner

Cold Spring Harbor Laboratory Press: 2006. 107 pp. \$19.95

Matt Ridley

When the history of genome sequencing at the turn of the millennium is written, it will centre on two battles. The battle over the sequencing of the human genome was bigger and more bitter than the one over sequencing the fruitfly genome; if the human genome was Waterloo, *Drosophila* was barely even Ligny or Quatre Bras. So why would anybody trouble to read, let alone write, a book about the lesser battle?

Michael Ashburner's answer is simple: for the story, not the history. He has written — or rather, he wrote, for these are his immediate reactions, mostly committed to paper at the time — an idiosyncratic, gonzo romp through the crazy days of 1998–99. His purpose, he writes, “frankly, was therapy”. This is not like General John Sulston's biographical justification of the genome campaign, *The Common Thread*, written with Georgina Ferry (National Academies, 2002), or James Shreeve's compendious war report from the camp of Marshal Craig Venter, *The Genome War* (Alfred A. Knopf, 2004). Rather, this is a field diary from a colonel in the infantry.

The story is by now a familiar one. In 1998 Venter burst into the human genome project, promising to produce a sequence quickly, privately and commercially, thanks to new sequencing machines and the shotgun technique he had used on bacteria. He announced that he would test his method first on fruitflies, winning the cooperation of Gerry Rubin, hitherto the chief sequencer of *Drosophila*, by promising the immediate release of data. Many arguments later, in November 1999, a flock of scientists gathered to witness an eleven-day ‘jamboree’, organized mainly by Ashburner, at the Maryland headquarters of Venter's Celera, to find, understand and count the genes in the finished fruitfly sequence.

Ashburner's book covers the year-and-a-half between these two events. It is a time of constant travel for the author: Cold Spring Harbor,

Crete, Florida, Heidelberg, Washington (twice), Iceland (a bird-watching holiday), Heidelberg again, Florida again, Washington, Zurich, Maine, Bloomington, Washington again... I may have missed a few. Being a scientist, Ashburner hates hotels (especially Marriotts), Microsoft, bad coffee and suits — the ones who negotiate on behalf of Celera. He likes or needs sushi, espresso, Lewis Carroll, beer and bouts of bird-watching.

The book rattles along with immediate, chaotic, rambunctious prose, digressing several times on every page into chatty and irreverent footnotes to explain who people are, how and why to find a certain restaurant, or where an aphorism comes from (‘Box and Cox’ is from an Arthur Sullivan operetta, apparently). Ashburner says that his education in the works of Geoffrey Chaucer taught him the value of footnotes.

The wheels came off the sequencing project on 1 November 1999, when Ashburner was frantically trying to finish several tasks before leaving for the jamboree. On that day, Celera released the fly genome to the NCBI's GenBank database, but with restricted access — to see any sequence, the user had to agree not to copy, re-sell or distribute it. This was not the open access that Venter had promised Rubin the year before. Venter was in a bind, desperate to show his backers that sequence data could have commercial value, and anxious to prevent rivals such as Incyte from selling Celera's data to others. So Venter tried it on. “The best way to find out if he was crossing the line,” wrote Shreeve in his history of the event, “was to stick out a toe and see if any alarms went off.”

This moment is a little hinge in history. Venter might have got away with it. Rubin was travelling. The NCBI thought this was what had been agreed. It was Ashburner in Cambridge who exploded. His expletive-ridden e-mails

jamboree, Ashburner persuaded Venter to pose in headphones like a NASA flight director in Celera's mission-control room.

The history of genome sequencing drives home the message that science is usually the daughter, not the mother, of technology. Sequencing the human, fruitfly and all other genomes was made possible by new machinery. The resulting information was a glorious and far-reaching addition to human knowledge, but only a distant harbinger of new commercial applications. But at least it has generated some enjoyable literature, of which this is a good example.

Matt Ridley is at Blagdon, Seaton Burn, Newcastle upon Tyne NE13 6DD, UK. His biography of Francis Crick will be published in June. www.mattridley.co.uk



Human frailties

The First Human: The Race to Discover Our Earliest Ancestors

by Ann Gibbons

Doubleday: 2006. 336 pp. \$26

Ian Tattersall

Once upon a time it was the 'missing link'. Then it was the earliest member of our genus *Homo*. Now the target of palaeoanthropology is the 'first human', our most ancient fossil ancestor not shared with the great apes. Such shifts in fashion demonstrate that, like any other branch of science, palaeoanthropology changes with the times. Some things, however, have remained comfortably constant. One of these is the regular public laundering of palaeoanthropology's dirty linen, which has long been a mainstay of the popular science literature, combining as it does the delights of gossip and voyeurism with a fascinating subject matter. And as a few years have passed since Jon Kalb's entertaining and instructive *Adventures in the Bone Trade* (Copernicus, 2000; reviewed in *Nature* **410**, 517–518; 2001), perhaps it is indeed time to trawl again through the rich and swirling palaeoanthropological pond, as Ann Gibbons does in her absorbing new book, *The First Human*.

Little more than a decade ago, the human story seemed to begin essentially with *Australopithecus afarensis*, the early hominid species from eastern Africa epitomized by Lucy, the most famous hominid fossil of them all. With representatives from a variety of sites spanning the period between about 4 million and 3 million years ago, the small-brained and diminutive *A. afarensis* was an upright biped when on the ground (albeit in a manner somewhat different from ours), but retained enough ancestral features to be viewed today as a bipedal ape.

The key feature that had set humankind on its unique evolutionary path had been energetically debated for a century, but the combination of characteristics displayed by Lucy and her kind seemed to settle the argument in favour of two-legged terrestrial walking. This conclusion was bolstered by palaeo-environmental data, which indicated that, over the past several million years, Africa has been subject to a drying trend, leading to forest being replaced by woodland and grasslands. In some areas at least, this forced hominoid populations to the ground, as their ancestral forest habitat disappeared.

Over the past dozen years, a slew of even more ancient hominid finds has been claimed from various parts of Africa. First there was *Australopithecus* (later *Ardipithecus*) *ramidus*, published in *Nature* in 1994. This fragmentary form is 4.4 million years old, and was later joined by bits and pieces of a claimed relative up to 5.8 million years old. In 1995 there followed *Australopithecus anamensis*, from

Kenyan sites up to about 4.2 million years old. In the 'really ancient' category (molecular estimates of the divergence of the human and ape lineages run at around 6 million to 8 million years ago), *Orrorin tugenensis* was described from 6-million-year-old deposits in Kenya in 2001, followed a year later by *Sahelanthropus tchadensis*, from slightly older sediments in Chad.

Each of these poorly known species derives its putative hominid status essentially from the claim that it was an erect biped. But upright locomotion is inferred from very indirect evidence in every case except *A. anamensis*, and these fossils make up an oddly assorted group. If they are all indeed descended from a unique common ancestor, then from the very beginning our precursors were experimenting wildly with different ways to be hominid. On the other hand, even if all these forms really were bipedal, perhaps, faced with similar environmental pressures, several hominoid lineages sharing a tendency to upright arboreal postures might independently have adopted this approach to a partly terrestrial life. Amid all this uncertainty, though, one thing is plain: it is in the interests of the palaeoanthropologists concerned that 'their' fossils be hominid rather than merely representatives of now-extinct lineages of apes.

In such a situation it is only to be expected that there should be keen competition among the various groups of scientists involved in the quest for the first human. And so it is that, after a historical overview, Gibbons' book is largely devoted to recounting the sometimes savage rivalries between these factions over the past decade and a half. Virtually all the protagonists are smart, dedicated and hardworking scientists, but some are also worthy of Machiavelli's *The Prince*. If their energies were devoted simply to demonstrating the hominid status of their fossils by any means available, the worst that could result would be a certain measure of confusion. Which is what we have: after years of argument, we still have no idea what the first human ought to look like. But to complicate matters further, palaeoanthropology is too often seen as a sort of zero-sum game, with the result that published specimens are often vigorously defended from the eyes of rival

scientists, making it impossible to test the describers' published assertions. What other branch of science keeps its primary data secret?

The ideal enquiry into this murky world might thus usefully ask some indiscreet questions. But even though it would be unfair to call *The First Human* hagiography, Gibbons' tone remains generally deferential. This is hardly surprising because, as a correspondent for *Science* magazine, Gibbons still has to talk regularly to her subjects. But it does limit her freedom to explain the many bizarre episodes she recounts, or indeed the intellectual framework of the underlying quest. So although this readable book will bring you up to speed on



many of the sociocultural strands of the search for the first human, you might not put it down feeling that you have really gained a mature understanding of the issues involved in recognizing and interpreting our earliest ancestors. But this is emphatically not Gibbons' fault: you won't get it from reading the primary literature either. ■

Ian Tattersall is in the Division of Anthropology, the American Museum of Natural History, New York, New York 10024, USA.



Science's secret service

The Jasons: The Secret History of Science's Postwar Elite
by Ann Finkbeiner
Viking: 2006. 304 pp. \$27.95

Daniel S. Greenberg

Advice from science to politics runs along diverse paths. Einstein wrote to US President Franklin D. Roosevelt to warn of the possibility of nuclear explosives. C. P. Snow asserted that the physicist Frederick Lindemann's close friendship with Churchill gave him "more direct power than any scientist in history". In response to Sputnik, President Eisenhower established a committee of scientists to advise the White House. And nuclear physicist Edward Teller exalted the Star Wars defence programme to a receptive President Reagan. Renown, personal relationships and committees have all linked science to government in the modern era.

And then there is the group known as Jason, which is unlike any of them. Working exclusively for the US government, to which it is financially tethered, Jason consists mainly of science professors — mostly self-selected and working in secret — who devote a large part of each summer to intense, collaborative studies of worrisome scientific and technological issues confronting the government. The title Jason apparently has a playful origin, being derived from the myth of the Argonauts.

Since Jason was founded in 1960, about 100 scientists have been taken into its ranks, of whom 30 to 60 are active at any one time. Physicists predominated in the early days, as befitted the reigning scientists of the Cold War, but life

scientists and others have been included since. At the latest tally, 11 have been Nobel laureates and 43 were members of the US National Academy of Sciences. Among them are some of the great names of science: Freeman Dyson, Richard Garwin, Charles Townes, Marvin Goldberger and John Wheeler. Most of Jason's studies concern military matters, and about half to three-quarters of its reports are classified; the exact proportion is unknown, because Jason, although not itself a secret, generally shuns publicity.

The public record contains little mention of Jason, with just a few published accounts, largely suspicious of the secretive proceedings of superstar professors temporarily employed by the Pentagon on unknown projects. But the veil is partly lifted by Ann Finkbeiner, who teaches science writing at Johns Hopkins University; her book *The Jasons* is based on interviews with 36 of the group's members.

Finkbeiner describes her book as "less a respectable history than a series of stories in more or less chronological order". She tells of Jason's proposal to dampen the fury of the Vietnam War by building a high-tech barrier of sensors and mines, backed by artillery and aircraft, to throttle the passage of North Vietnamese troops and supplies along the Ho Chi Minh trail. Embraced by the beleaguered defence secretary, Robert McNamara, the barrier lacked an important ingredient: support from the military, which was distressed by its costs and echoes of Maginot Line immobility. Although a stretch of the barrier was built, a Pentagon official retrospectively labelled it "horribly naive". The episode was not fruitless,

however, according to several Jasons, who claimed the barrier as the prototype for the electronic battlefield.

Involvement in the Vietnam War, even if aimed at lessening its destructiveness, brought self-doubt and opprobrium to the Jasons. Student activists denounced and harassed them after stolen records were depicted as implicating Jason deeply in the conduct of the war. A 1967 Jason study, "Tactical Weapons in South-east Asia", demonstrated the futility of nuclear weaponry in the Vietnam War but remained classified for about 20 years. "Jason became the devil," a Pentagon official remarked. "No matter that some of them had excellent doveish credentials, they were Jason."

The Jasons is particularly valuable in its illumination of the realpolitik of scientific advice. Claire Max, one of the few women Jasons, remarks: "I get morally indignant about something that the government is doing that doesn't make technical sense, and I suggest to the rest of my colleagues at Jason that we do a summer study. And they tactfully point out that there's no point in doing it if nobody's going to listen."

Sidney Drell warns not to expect too much when providing scientific advice to government. "The ratio of output to input in doing government work is never high," he observes. "I take great comfort in the fact that for fifty-seven years we've managed severe crises and not used nuclear weapons. I just want to see that we keep our wits about us. And I haven't lost my optimism."

Among these stars of the scientific establishment, attitudes vary about going public with the weighty issues of their government work. Richard Garwin, the quintessential insider, is legendary for speaking out publicly on topics ranging from the environmental impact of supersonic flight to the control of nuclear arms. Some half-a-dozen current or past Jasons are prominent in the leadership of the Federation of American Scientists, which pioneered scientific activism over arms control and international collaboration. In contrast, Jason member Rich Muller argues: "Once you come out publicly for something, you're basically taking sides, and your value as an adviser is decreased."

Throughout its nearly 50-year history, Jason's free-wheeling style has often grated on the defence officials who pay its bills. Clashes have ensued, and Jason has had its ties to the Pentagon rearranged and its client base expanded to other government agencies. With its intolerance of dissent, the George W. Bush administration has been especially trying for Jason, particularly on the sensitive issue of who is appointed. However, Jason goes on, although its traditional cohesiveness is challenged by the growing interdisciplinarity of science, and family and work patterns that make it difficult for members to get away for six weeks in the summer.

Elite academic scientists working on secret government projects would seem to be the

basis for a great story. But *The Jasons* rings few dramatic bells. And the author observes: "Jasons think they've done the country good, but they haven't got much clear evidence."

Perhaps a more sceptical attitude towards the Jasons in particular and the advisory process in general would have paid off. My own observation is that politicians seek scientific advice to support their preconceptions, not to steer them to wisdom. But where secrecy previously obscured public view of the Jasons, reverence is the problem here. Stockholm

syndrome is not confined to hostages who become enamoured of their captors: it can also afflict observers who fall for their subjects. As Finkbeiner wrote of an anonymous Jason identified only as "Dr. Y": "She might look plain if she didn't move so gracefully, and if she didn't sit so quietly, self-possessed and intense and honest to the bone."

Daniel S. Greenberg is a journalist in Washington. His new book, *Science for Sale*, will be published next year by the University of Chicago Press.

difference in complexity and the likelihood of emergent properties, it is unlikely that probing the ganglia of *Aplysia* can tell us how our brain recalls our first love or our father's face.

Nevertheless, the reductionist approach to behavioural plasticity, of which Kandel's work is the prime example, is a success story that has given us models of molecular and cellular plasticity that propose how experience affects nerve cells. It is now for those who follow Kandel to link the molecular and cellular level with the systems level of analysis. This integration is the major challenge facing the science of memory, and might require, in addition to new methodologies, a change of *zeitgeist* or an amalgamation of approaches.

Kandel's book is enthusiastically recommended as a captivating account of the career of a prominent leader in contemporary neuroscience. The author is not only an authoritative scholar but also a marvellous popularizer and narrator, who brings to the story an attractive mix of facts, personal touches and wisdom, seasoned with reflective humour. But *In Search of Memory* is not just about science: it is also about history and identity. Kandel is a devoted scientist, humanist, family man and proud Jew. He follows, by his own definition, the "Talmudic tradition writ large. But rather than annotate a religious text, we annotate texts written by evolutionary processes working over hundreds of millions of years."

Kandel was just a child when he emigrated from Austria to the United States, but the Holocaust and the trauma of European Jewry are deeply embedded in his memory. His contempt for racism is clear. In the background hovers the terrible awareness that many of his generation perished, unknown, in concentration camps before they had a chance to explore and contribute to the world. When the Austrian president contacts Kandel and expresses his desire to honour the new Nobel laureate of Viennese origin, Kandel's reaction is to organize a symposium in

Vienna to acknowledge Austria's central role in the Nazi atrocities and evaluate the significance for scholarship of the disappearance of the Jewish community in Vienna. Kandel's accounts of incidents during this visit to Vienna should be read carefully by those who ignore lingering undercurrents of anti-Semitism.

The Greek poet Constantine Cavafy, in his poem *Ithaka*, which recounts the return of Odysseus to his homeland, advises his hero not to hasten:

*As you set out for Ithaka,
Hope the voyage is long,
Full of adventure, full of discovery...
Better if it lasts for years...
Ithaka gave you the marvellous
journey
Without her you would not have
sailed away.*

A journey to remember

In Search of Memory: The Emergence of a New Science of Mind

by Eric R. Kandel

W. W. Norton: 2006. 510 pp. \$29.95

Yadin Dudai

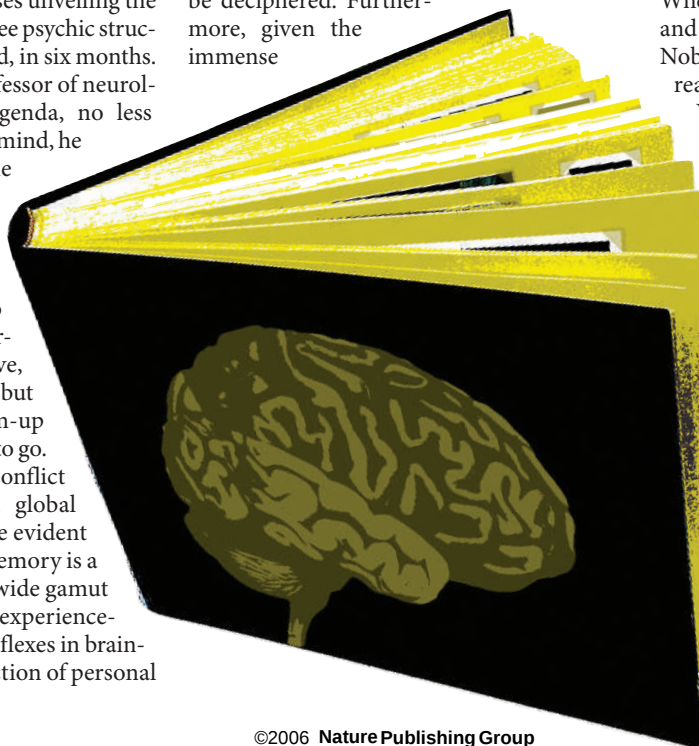
Few can interlace their autobiography with the evolution of a scientific paradigm. Even fewer can weave such a story seamlessly. Eric Kandel is one of these. His career, from his training in Harry Grundfest's laboratory at Columbia University in New York more than fifty years ago to a remarkably productive present, also at Columbia, epitomizes his ardent reductionist approach to the neural sciences. Its formal pinnacle was the Nobel Prize in Physiology or Medicine, which Kandel shared in 2000 with Arvid Carlsson and Paul Greengard for their discoveries concerning signal transduction in the nervous system.

Kandel's intellectual journey in neuroscience can be traced back to his first encounter with Grundfest. The enthusiastic medical student, with a strong background in history and literature, proposes unveiling the brain substrates of Freud's three psychic structures, the ego, superego and id, in six months. Grundfest, the seasoned professor of neurology, suggests a different agenda, no less grandiose: to understand the mind, he replies, we need to look at the brain one cell at a time. The narrative of the brain sciences in the past century is made up of attempts to negotiate between these two extremes. In his admirable personal version of this narrative, Kandel is still a grundfestian, but appreciates that the bottom-up approach still has a long way to go.

Nowhere, perhaps, is this conflict between reductionism and global approaches to the brain more evident than in memory research. Memory is a term applied nowadays to a wide gamut of functions, ranging from experience-dependent modification of reflexes in brainless organisms to the recollection of personal

events in their investigators. Even if such an inclusive definition is accepted, the distinction should still be made between memory as a process and as an item with mental content.

The process is assumed to be subserved by the plasticity of synapses, the functional contacts between nerve cells. One could further assume that the basic building blocks of the plasticity machinery were conserved in evolution. If this is the case, why not approach memory by studying its simplest forms in the simplest of organisms? This philosophy has guided Kandel and given the timid sea-slug *Aplysia* a prominent position in textbooks of neuroscience. The approach, anchored in the achievements of molecular biology, has proved highly productive in identifying plasticity mechanisms that subserve memory. What it doesn't address satisfactorily is the content and meaning of memory items. This requires an understanding of how brains encode specific pieces of mental information. Many would argue that this calls for the spatiotemporal codes of neuronal populations to be deciphered. Furthermore, given the immense



The reward, according to Cavafy, is the journey, rather than the goal. Both the young Kandel who met Grundfest and the mature, imaginative investigator of the *Aplysia* epoch seem to have valued the goal at least as much as the journey. But reading these memoirs, one senses that, over the years, Kandel's appreciation of the journey itself has increased. Is Ithaca attainable for those who study memory?

If the goal is to chart and analyse plasticity in neuronal terrain in fine detail, then the kandelian *Aplysia* paradigm is a tremendous leap forwards. If it is to understand how recollecting humans think, feel and plan, we might need more Kandels. ■

Yadin Dudai is in the Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel.

concerns with biodiversity loss and his contributions to the development of the fledgling field of environmental ethics.

Each article is preceded by a brief essay in which Wilson explains its context and outlines more recent developments in the field. Although not as thorough and telling as those by W. D. Hamilton in *Narrow Roads of Gene Land*, vols 1–3 (Oxford University Press, 1997–2005), these brief essays are interesting and informative about some essentials of Wilson's personality. For example, Wilson explains that he revised the ant genus *Pheidole* as “a hobby, a form of relaxation” during which he “listened to classical and soft rock music...at odd hours” in his home laboratory. The outcome was *Pheidole in the New World* (Harvard University Press, 2003), an 800-page monograph with more than 5,000 drawings and a list of 624 species, of which 337 are new to science (see *Nature* 424, 727; 2003) — an achievement beyond what most scientists would ever dream of accomplishing in a lifetime.

Nature Revealed contains all 60 articles in their original form. Unfortunately, the size reduction of several articles resulted in print so small as to be reminiscent of the labels on insect museum specimens. The diversity and technical nature of some of the articles might make this book difficult to read from cover to cover, but it remains a treasure for those interested in science and history. The progression of articles highlights the path of Wilson's journey across different disciplines in an attempt to bridge gaps between them, including the gulf between the humanities and the sciences. These are important messages in this age of ultraspecialization and disciplinary compartmentalization. ■

Laurent Keller is in the Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne 1015, Switzerland.

The ant trail

Nature Revealed: Selected Writings, 1949–2006

by Edward O. Wilson

Johns Hopkins University Press: 2006. 736 pp. \$35

Laurent Keller

Edward O. Wilson is one of the most distinguished scientists and thinkers of our time. He is best known for his twenty or so books, exploring topics as diverse as sociobiology, human nature, ant taxonomy, biodiversity and the philosophy of knowledge, but he has also published a large number of scientific articles and other works. *Nature Revealed*, a compilation of 60 of these articles, illustrates how his interest has ranged between fields as diverse as entomology, ethology and philosophy.

Wilson's first publication was in 1949 in a local journal called *Alabama Conservation*. In this paper, Wilson, then a 19-year-old senior at the University of Alabama, reports the distribution of what was then called the imported fire ant. His second article was published in *Evolution*, an influential scientific journal. In this article Wilson reports that there are two phenotypic variants of the imported fire ant. By switching queens from one colour-type colony to the other, he proved a hereditary basis for the phenotypic differences observed and proposed several explanations for how one morph replaced the other. Incidentally, later studies revealed that the two colour morphs are actually two distinct species, now called the black imported fire ant (*Solenopsis richteri*) and the red imported fire ant (*Solenopsis invicta*), the latter of which has become one of the worst invasive pest species in the world.

The difference in the content of these two articles foreshadows Wilson's main gift: his ability to be hugely integrative and make bold syntheses. In the scant two years that separated the publication of these two reports, Wilson had vastly broadened his interests, allowing him to interpret his field observations within a solid framework of evolutionary biology and population genetics.

The progression of articles in *Nature Revealed* demonstrates, again and again, Wilson's endless capacity to put scientific findings into a broader context and to bridge gaps between disciplines. For example, between

1959 and 1962 he published a few specialized papers detailing the nature of communication within ant societies. By 1963 he had realized that pheromones were important, and had ventured to propose general principles in a *Scientific American* article entitled simply “Pheromones”. In this piece, Wilson also speculated for the first time on the existence of human pheromones. On the basis of data by French biologist J. LeMagnen, who showed that the odour of exaltolide can only be perceived clearly by sexually mature women at the time of ovulation, Wilson stated that although these observations “hardly represent a case for the existence of human pheromones...they do suggest that the relation of odours to human physiology can bear further examinations”. Such examinations did indeed reveal that pheromones are implicated in some human behaviours.

In the same vein, the sequence of publications reveals the steps that led Wilson from reports of patchy ant distributions in the rainforests of New Guinea to the synthesis with Robert MacArthur of the influential theory of island biogeography. Similarly, the content of several articles demonstrates Wilson's growing



A cross-cultural relationship

The Literary Animal: Evolution and the Nature of Narrative

edited by Jonathan Gottschall & David Sloan Wilson

Northwestern University Press: 2005.
375 pp. \$79.95 (hbk), \$29.95 (pbk)

Rebecca Goldstein

Like certain partners in long-term marriages, the arts and sciences have often got along best by agreeing to ignore one another. When they do engage, uncomplimentary comparisons are hurled, leaving one side or the other, or more likely both, feeling misunderstood and unappreciated, unable to do better than lash out in impotent resentment: "You'll never understand me."

But lately these two have been getting on rather better. A significant number of novelists, playwrights and film-makers have turned to the sciences for themes and characters, exploring questions ranging from quantum mechanics to neural networks in ways that not only do justice to the ideas, but also to the passions of those who pursue them. Likewise, a small cadre of scientists have begun to take the arts seriously, looking to them for important questions and data. Such heroic attempts to breach the gap in communication between the two progenitors of our culture are certainly an encouraging sign.

It has been primarily through the emerging field of evolutionary psychology that the arts have come to the respectful attention of the sciences. Itself undergoing an evolution from what had sometimes seemed to be 'just so' hypotheses to a refined theory with explanatory muscle, evolutionary psychology offers the means

of bringing together the arts and sciences in a mutually reaffirming relationship: each gives the other a means of gaining insight into itself, with neither feeling slighted or outsmarted.

The promise of the evolutionary approach is admirably demonstrated in *The Literary Animal*, a collection of essays penned by contributors from both sides of the divide. The book provides the exhilarating sense of witnessing something new taking shape on the intellectual horizon, a 'proto-paradigm' that, although grounded, is not yet set in stone, with big questions still being debated.

The Literary Animal focuses on just the narrative arts, novels and drama, a sub-genre of the meeting of the arts and sciences that has been dubbed 'literary darwinism'. One of the basic questions is whether literature, of all the art forms, presents a unique set of issues for evolutionary psychology. The very existence of this book implies that it does, although it is not obvious that all the contributors would agree.

The attention that evolutionary psychology pays to literature is grouped into three areas, all addressed, albeit unequally, in *The Literary Animal*. The first is the evolutionary problem posed by the arts. The puzzle, as the book's editors Jonathan Gottschall and David Sloan Wilson, put it, is this: "In ancestral environments characterized by intense competition for survival and reproduction, how could the evolutionary process 'allow' any animal to spend (waste?) so much time producing, elaborating, and consuming art — time that could be spent pursuing mates and other quarry?"

An issue that divides evolutionary psychologists is whether the arts are adaptations to the biological exigencies of surviving and reproducing, or by-products of adaptations that evolved to perform other functions. This question attracts its fair share of debate in the book, with some contributors arguing vociferously for the 'art is adaptation' point of view, although each offers a different account of what the artistic adaptation is designed to do: perhaps increase our ability to sustain attention (Brian Boyd), or regulate "the human cognitive behavioural system" (Joseph Carroll). Michelle Scalise Sugiyama argues that the narrative arts have an adaptive purpose all of their own, unshared by the other art forms. The fact that narrative is set forth in language adapts it to be an "information storage and transmission system" that works by simulating — and stimulating — experience.

A second reason why evolutionary psychology pays attention to literature is as a source of evidence, buttressing its claims about human nature, including the claim that there is such

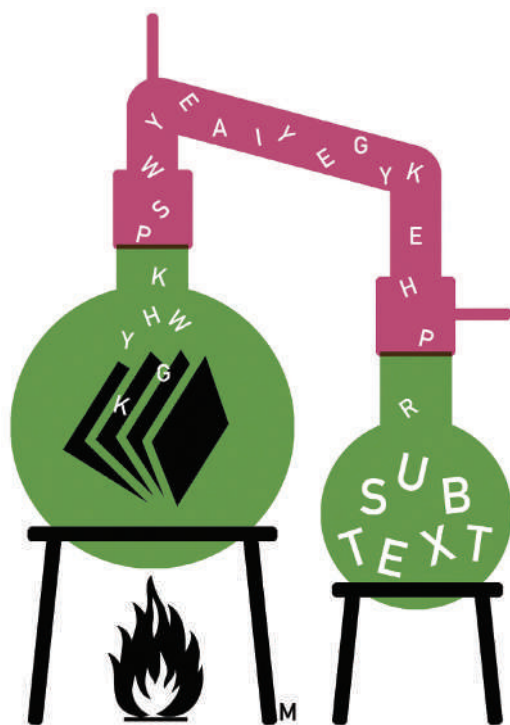
a thing at all. The novelist Ian McEwan's contribution argues this claim: "It must be clear by now, I think, that the exercise of imagination and ingenuity as expressed in literature supports Darwin's view. It would not be possible to read and enjoy literature from a time remote from our own, or from a culture that was profoundly different from our own, unless we shared some common emotional ground, some deep reservoir of assumptions, with the writer."

Illustrating the search for literary data supporting the existence of a universal human nature are essays considering the importance of the male-male bond in epics and romances, the theme in both Japanese and Western literature of men rejecting children whom their wives have conceived in adultery, and a study that connects the two sorts of male hero presented in novels — the dark and dangerous Byronic 'cad', and the safe and reliable 'dad' — to female choices for short- and long-term pair bondings.

The third cluster of questions is the most tantalizing: can literary darwinism be transformed into a responsive literary theory that enhances our understanding and enjoyment of specific literary texts? A work of art, almost by definition, demands and rewards disciplined attention, and humanities scholars have long devoted themselves to techniques aimed at coaxing forth the full meaning of a work. The theories marshalled towards this end all have in common the claim that, despite the individual face that each work bears, there are general theoretical abstractions that can be applied to each that will reveal what is really going on within the text.

Although this third area seems closest to the meaning of the term 'literary darwinism', it receives the least attention in *The Literary Animal*. Only Carroll's chapter on "human nature and literary meaning" provides a close literary analysis of a specific text, Jane Austen's *Pride and Prejudice*. The result is certainly worthy; his reading of Austen is subtle and supple. But the depth of his analysis derives more from his purely literary observations on the multiplicity of points of view involved in literary works than from the general darwinian points he has to make.

It remains to be seen whether evolutionary psychology can spawn a true literary theory that helps us understand narrative in general — why our species pays it so much attention, and how its universal themes reflect facets of human nature explained by evolutionary psychology — and whether it will also yield a method subtle and flexible enough to deepen our understanding of specific literary works. In the meantime, *The Literary Animal* provides a fine example of the riches that await us as the arts and sciences stop competing and start communicating. ■
Rebecca Goldstein is a philosopher and writer. Her most recent book is *Betraying Spinoza: The Renegade Jew Who Gave Us Modernity*.



NEWS & VIEWS

VACCINES

Engineering immune evasion

John R. Mascola

One obstacle to realizing the promise of viral vectors for vaccine delivery is pre-existing immunity to such vectors. An adroit application of structure-based design points to a way around that problem.

There are still no vaccines against such devastating and widespread diseases as malaria, tuberculosis and AIDS. Because the traditional approach of live-attenuated vaccination is not feasible for most diseases, scientists have turned to molecularly engineered viruses that contain pathogen-specific gene inserts. Such viral vectors direct host cells to produce the foreign protein of interest, thus prompting a pre-emptive immune response. Among the most promising viral vectors is a form of common-cold virus known as adenovirus serotype 5. The recombinant adenovirus vectors (rAd5) cannot replicate and can be safely administered, and they elicit both of the two main types of immune response — secreted antibodies and disease-fighting T cells.

There is a problem, however. The existing immunity to rAd5 in many adults means that the vector could be neutralized before it can have an effect. Hence the work of Roberts *et al.*¹, described on page 239 of this issue. They have taken a fresh approach to the molecular engineering of rAd5, one that has the potential to circumvent anti-vector immunity and expand the applicability of such vectors for human vaccination.

Numerous viral vectors are being studied for use in gene-based vaccine strategies. The most commonly used vectors are derived from poxviruses, alphaviruses and adenoviruses. Among these, rAd5 is the best characterized and is perhaps the most attractive for vaccine development. As a stand-alone vaccine, rAd5 can elicit different types of T-cell immunity (those due to CD4 and CD8 cells), and more potent immune responses can be achieved with a 'prime-boost' approach. For example, use of vehicles known as DNA plasmids followed by boosting with rAd5 can generate durable antibody and T-cell immune responses^{2,3}. Preclinical studies of rAd5 vaccines include vaccines against Ebola, SARS, HIV-1 and anthrax^{4–7}, and phase II human clinical studies of rAd5 HIV-1 vectors are in progress.

But anti-vector immunity may be a serious limitation. Adenovirus serotype 5 is common — depending on the geographical region of the world, most adults are exposed to it and

develop some level of immunity. This may lessen the effectiveness of rAd5 as a vaccine vector. Potential ways around this problem include the use of adenoviruses derived from other human serotypes or from non-human animal species. Indeed, there are more than 50 known human adenoviral serotypes, some of which are quite rare in the human population. The genetic manipulation required to engineer alternative serotypes is not trivial, however. The rAd5 vectors contain specific genetic deletions that render them unable to replicate. This contributes to their safety, but also means that specially engineered cells must be used to produce them. The advantages of rAd5 are that the necessary groundwork has been laid, in terms

They have taken a fresh approach to the molecular engineering of rAd5, one that has the potential to circumvent anti-vector immunity and expand the applicability of such vectors for human vaccination.

of basic molecular engineering and production of the vector, and that it has been through the regulatory approval process for use in humans. Adaptation of other serotypes will require a methodical process of research and development, and safety testing. Furthermore, preliminary data^{5,8} from other serotypes, such as rAd35 and rAd11, suggest that they may be less immunogenic — that is, less effective in producing immunity — than rAd5.

With this as background, Roberts and colleagues¹ took advantage of our improved understanding of anti-vector immunity, coupled with structural data about viral proteins, to derive a rational approach to re-engineering the rAd5 vector. In adenoviruses, the viral DNA is surrounded by a protein shell called a capsid that contains hexon and penton subunits. Because host antibodies that neutralize rAd5 are directed against the hexon subunit, Roberts *et al.* studied the atomic structure of

this protein to understand where antibodies would probably bind. Molecular modelling revealed that the seven hypervariable regions (HVRs) of the hexon form the outer surface of the protein, making the HVRs a likely target for antibody binding.

By exchanging all seven HVRs of rAd5 with those of the rare adenovirus serotype 48 (Ad48), the authors constructed a chimaeric adenovirus that could potentially evade the neutralizing antibody response against rAd5. The core structure of the hexon protein was not altered, so the resulting HVR-chimaeric rAd5 vectors retained their ability to grow well in culture and, importantly, the immunogenicity of the chimaeras was comparable to that of rAd5. As Roberts *et al.* hoped, when the HVR chimaeras were administered to mice or monkeys that had antibody immunity to rAd5, there was no decrease in the immunogenicity of the vector.

These data provide a proof-of-concept that viral vaccine vectors can be engineered to evade pre-existing immunity. The results are a tribute to the application of modern immunology and structural biology to vaccine design. The potential of this technology is considerable. One can envisage the construction of numerous HVR chimaeras that could be used to vaccinate against various pathogens. Thus, if rAd5 itself were used to vaccinate children against malaria, a chimaeric vector could still be used as an HIV vaccine. Furthermore, the use of multiple chimaeric serotypes could allow booster vaccinations to sustain the long-term immune memory response needed for durable immunity.

Yet we are still some time away from studies in humans. Vaccine developers will have to show that these new HVR-chimaeric rAd5 vectors can be manufactured, and that they have stable gene inserts, can pass regulatory review and, finally, are immunogenic in humans with pre-existing immunity. Current rAd5 vectors for HIV-1 are being evaluated in phase II human trials that will more precisely define the extent and effect of pre-existing anti-vector immunity. As we await these data, chimaeric vectors can be manufactured

and tested in humans, so that we can further assess the potential effects of anti-vector immunity.

John R. Mascola is at the Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 40 Convent Drive, Bethesda, Maryland 20892-3005, USA.
e-mail: jmascola@nih.gov

1. Roberts, D. M. *et al. Nature* **441**, 239–243 (2006).
2. Casimiro, D. R. *et al. J. Virol.* **77**, 7663–7668 (2003).
3. Santra, S. *et al. J. Virol.* **79**, 6516–6522 (2005).
4. Bangari, D. S. & Mittal, S. K. *Vaccine* **24**, 849–862 (2006).
5. Shiver, J. W. & Emini, E. A. *Annu. Rev. Med.* **55**, 355–372 (2004).
6. Barouch, D. H. & Nabel, G. J. *Hum. Gene Ther.* **16**, 149–156 (2005).
7. Sullivan, N. J. *et al. Nature* **424**, 681–684 (2003).
8. Lemckert, A. A. *et al. J. Virol.* **79**, 9694–9701 (2005).

SOLAR SYSTEM

Interplanetary kidnap

Alessandro Morbidelli

Triton, Neptune's largest moon, was probably part of a two-body object similar to the Pluto–Charon system. This tandem might have been ripped apart when it strayed too close to the planet that Triton is now orbiting.

The neptunian moon Triton weighs in at 1.4 times the mass of Pluto, making it the largest irregularly orbiting satellite in the Solar System. So how did this kept giant come to be where it is? On page 192 of this issue¹, Agnor and Hamilton advance a capture mechanism that, if correct, could have repercussions for the life stories of other, similar moons.

A cohort of satellites surrounds all four giant planets in the Solar System — Jupiter, Saturn, Uranus and Neptune. These satellites are divided into two distinct groups, regular and irregular, according to their orbital characteristics. Regular satellites are closer to their parent planet, with orbits that are essentially circular and that lie on the planet's equatorial plane. These satellites thus constitute miniature solar systems around their planet. And just as the planets of the Solar System are thought to have formed from a disk of gas and dust (the protoplanetary disk) orbiting the Sun, so the regular satellites are assumed to have formed from a 'planetesimal disk' orbiting their planet.

Irregular satellites, in contrast, are more distant from their planet and typically have orbits with larger eccentricities (a measure of deviation from a perfect circle) and/or inclinations. About half the irregular satellites orbit their planet in the retrograde direction; that is, in the opposite direction to the rotation of their planet. Because of these strange orbital characteristics, the general assumption is that these satellites formed on heliocentric orbits, and only later were captured on elliptic orbits around giant planets.

Several mechanisms have been proposed for this capture process. Some invoke the effect of gas-drag exerted by the atmospheres of the planets, which were more extensive when the planets formed some 4.5 billion years ago than they are now, owing to the heat generated by the accretion process. Others posit the abrupt growth of a planet's mass (the 'pull-down

mechanism') as the culprit. Still others employ gravitational interactions or collisions with the system of regular satellites already established around the planet, or fortuitous encounters with other planetesimals on heliocentric orbits as they passed through the sphere of influence of the planet.

But none of these mechanisms seem appropriate for Triton, Neptune's huge retrograde companion. Its mass — in another word, its inertia — means it was unlikely to have been captured by interactions with the existing satellites or other passing planetesimals. Additionally, Neptune is assumed to have undergone slow growth, and never to have had an extended atmosphere; both the pull-down and gas-drag mechanisms would therefore have been inefficient.

Agnor and Hamilton postulate¹ that Triton was originally part of a binary object, for instance similar to the Pluto–Charon system. They show that if this binary had passed sufficiently close to Neptune at low velocity, the different forces acting on its two constituent bodies would have ripped them apart. Each of the constituents would have, relative to Neptune, a velocity that was essentially the vector sum of the velocity of the binary's barycentre (its centre of gravity) and its own orbital velocity relative to this barycentre (Fig. 1). Most of the time, the orbital motion of one of the bodies is opposed to the barycentre's motion, so the net velocity of this body relative to Neptune could easily have been smaller than the escape velocity from Neptune's gravitational field. Thus, it would have become captured in a bound, planetocentric trajectory.

To be a likely explanation of Triton's capture, this model requires that two conditions be met. First, the protoplanetary disk in which Neptune evolved must have contained a very large number of Pluto-sized objects². This condition cannot be checked directly, but is plausible: the protoplanetary disk is presumed to have had

a total mass of about 50 Earth masses^{3,4}, or 5,000 times greater than that of the Kuiper belt⁵. (This relic of the protoplanetary disk⁶, where Pluto resides, now orbits the Sun beyond Neptune.) Second, a substantial fraction of the large objects in the protoplanetary disk must have been binary. The likelihood of this is increased by the observation that between 10 and 15% of the objects in the Kuiper belt are two-bodied⁷. In addition, three of the four largest Kuiper-belt objects — in decreasing order of size, 2003 UB₃₁₃, Pluto and 2003 EL₆₁ — have satellites⁸. (The third largest, 2005 FY₉, is the odd one out.)

Although Agnor and Hamilton focus exclusively on Triton, it is tempting to conjecture that this mechanism applies to the capture of most of the irregular satellites. It has been pointed out that for all four giant planets, the number of irregular satellites larger than a specific size⁹ is about the same. This fact argues against the gas-drag and pull-down mechanisms for their capture: because the flux of planetesimals through the giant planets' orbits was about the same⁷, both mechanisms should have been much more effective for Jupiter and Saturn, which grew rapidly as gas giants, than for Uranus and Neptune, which formed more slowly in a gas-starved environment¹⁰.

The only thing that the giant planets have in common is the size of their sphere of gravitational influence, or Hill sphere. This fact, together with the giant planets' similar number of irregular satellites, suggests that some sort of two-body interaction inside the Hill sphere played the dominant role in the capture of such satellites. Additional support for this picture

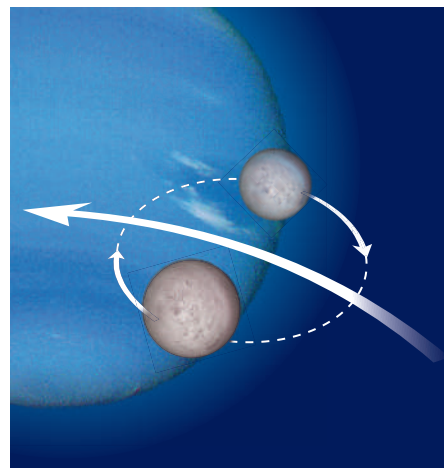


Figure 1 | Imminent capture. A two-body object approaches Neptune, the scenario envisaged by Agnor and Hamilton¹ for the capture of Triton. Both constituents of the binary rotate around the trajectory of their mutual centre of mass (barycentre), such that their own motion is almost half of the time with, and almost half of the time against, the motion of the binary as a whole. The net velocity of the constituents relative to the planet is accordingly increased or reduced; if the net velocity of a constituent drops below the escape velocity from a planet's gravitational field, it is captured, entering a new life as an irregular satellite of the planet.

comes from a model proposed by our group¹¹ of the origin of the Late Heavy Bombardment. This model implies that the irregular satellites were captured during this period between 4 billion and 3.8 billion years ago, which was characterized by a large number of collisions of asteroids and comets with the terrestrial planets. This was well after the disappearance of the gas and the growth of the planets.

The problem with two-body interactions is that the encounter in the vicinity of a planet of two planetesimals on independent heliocentric orbits is extremely improbable. The capture of irregular satellites from binary objects brilliantly circumvents this problem as, by definition, the two interacting bodies approach together.

I predict that the model proposed here for the capture of Triton will rapidly become a mainstay for models of the origin of irregular satellites, one of the principal open problems in planetary science. In general, irregular satellites are much less massive than Triton, and the capture mechanism requires that their orbital velocity inside the binary is large enough to

cancel out a substantial fraction of the velocity of the binary barycentre. Because this velocity increases with the total mass of the binary, these irregular satellites must originally have been secondary members of binaries with a large primary constituent. Such capture conditions, once thoroughly investigated, will unveil important constraints on the structure of the primordial protoplanetary disk. ■

Alessandro Morbidelli is at the Laboratoire Cassiopée, Observatoire de la Côte d'Azur, BP 4229, 06304 Nice Cedex 4, France.

e-mail: morby@obs-nice.fr

1. Agnor, C. B. & Hamilton, D. P. *Nature* **441**, 192–194 (2006).
2. Stern, S. A. *Icarus* **90**, 271–281 (1991).
3. Hahn, J. M. & Malhotra, R. *Astron. J.* **117**, 3041–3053 (1999).
4. Gomes, R. S., Morbidelli, A. & Levison, H. F. *Icarus* **170**, 492–507 (2004).
5. Bernstein, G. M. *et al.* *Astron. J.* **128**, 1364–1390 (2004).
6. Malhotra, R. *Astron. J.* **110**, 420–429 (1995).
7. Stephens, D. C. & Noll, K. S. *Astron. J.* **131**, 1142–1148 (2006).
8. Brown, M. E. *et al.* *Astrophys. J.* **639**, L43–L46 (2006).
9. Jewitt, D. & Sheppard, S. *Space Sci. Rev.* **116**, 441–455 (2005).
10. Pollack, J. B. *et al.* *Icarus* **124**, 62–85 (1996).
11. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. *Nature* **435**, 466–469 (2005).

CELL BIOLOGY

Cracking the calcium entry code

Anant B. Parekh

A sharp increase in the concentration of calcium ions in a cell is a key biological signal. Now a vital component of a major route by which calcium ions flow into cells has been identified.

Communication between and within cells is essential for the development and survival of any complex organism. Cells converse with each other mainly through a complement of chemical messengers, including neurotransmitters and hormones, that impinge on the cell surface, generating further signals (second messengers) within the cell that elicit the appropriate responses. Although several hundred hormones, neurotransmitters and other molecules can stimulate cells, the number of intracellular second-messenger systems they activate is remarkably small. Perhaps the most widespread and ubiquitous of the second messengers is the calcium ion. Two reports, one in this issue by Feske *et al.*¹ (page 179)¹ and one in *Science* by Vig *et al.*², uncover an essential component that allows cells to regulate their intracellular Ca²⁺ concentration.

A sharp rise in intracellular Ca²⁺ concentration can activate a disparate range of responses in almost all cells in the animal kingdom. Such rises stimulate neurotransmitter release, muscle contraction, cell metabolism, cell growth and proliferation, as well as processes culminating in cell death. Cells can increase their intracellular Ca²⁺ concentration in two ways: by releasing Ca²⁺ from intracellular stores, or by allowing

Ca²⁺ across the cell membrane. The Ca²⁺ stores have a limited capacity, so Ca²⁺ influx into the cell drives most of the cellular responses.

A primary route for Ca²⁺ influx is through 'store-operated channels' in the cell membrane³. The prototypic store-operated channel is the ubiquitous CRAC (for Ca²⁺ release-activated Ca²⁺) channel^{4,5}. Entry by this route is activated by a fall in Ca²⁺ within the endoplasmic reticulum⁶ — the labyrinthine network of membranes used to transport various substances around the cell. The endoplasmic reticulum usually holds a considerable stock of Ca²⁺ ions, without which it cannot function. However, a fall in Ca²⁺ concentration here, usually in response to the second messenger inositol 1,4,5-trisphosphate, somehow translates into the opening of store-operated Ca²⁺ channels. The molecular basis of this route of entry remains one of the more enduring mysteries in cell biology.

Once opened, CRAC channels enable Ca²⁺ ions to enter the cell. But despite their biological and clinical importance, very little is known about how these channels are activated, let alone their molecular composition. In breakthrough papers, Feske *et al.*¹ report that a protein they call Orai1 (encoded by

human gene *FLJ14466*) is an essential component of the CRAC channel complex, and Vig *et al.*² identify the same gene (although they call the encoded protein CRACM1) and reach the same conclusion.

The work of Feske *et al.*¹ built on a significant finding made last year that a protein called stromal interaction molecule (STIM1) is required for the activation of store-operated Ca²⁺ influx^{7,8}. STIM1 spans the endoplasmic reticular membrane and has a Ca²⁺-binding structural motif that may 'sense' the store's Ca²⁺ concentration. As the store loses Ca²⁺, STIM1, which is diffusely dispersed throughout the endoplasmic reticular membrane, becomes redistributed into discrete spots (puncta) in the cell periphery. Whether it remains just below the membrane or is inserted into it remains controversial.

However, although necessary, the presence of STIM1 is not enough to activate CRAC channels. Patients with severe combined immunodeficiency (SCID) have impaired CRAC channel activity in their T cells that renders these immune cells defective. But STIM1 in these cells is normal, nor does overexpression of STIM1 restore Ca²⁺ influx⁹. So a molecule downstream of STIM1 must be responsible for the SCID defect.

Feske *et al.*¹ demonstrate convincingly that this molecule is Orai1. To start with, they genetically mapped the mutation underlying SCID to a region on human chromosome 12. Because store-operated Ca²⁺ influx occurs in a similar manner in fruitflies and mammals, and because the fruitfly is the more easily manipulated in genetic experiments, the authors conducted a thorough search of the fruitfly genome for genes involved in store-operated Ca²⁺ influx. They found that the fruitfly gene encoding the Orai protein is a key component of the process. Crucially, a human relative of this gene, encoding Orai1, mapped to the same region on chromosome 12 as that linked to the SCID mutation. Elegant confirmation that Orai1 is responsible for the CRAC channel defect seen in SCID was provided when Feske *et al.* overexpressed Orai1 in SCID T cells: the normal protein rescued store-operated Ca²⁺ influx and CRAC channel activity, but a mutated version did not.

A similar search of the fruitfly genome by Vig *et al.*² identified the same gene as being central to CRAC channel activation. Reducing the expression of this gene abolished CRAC activity in three different cell types. So two independent groups have identified the same protein as being fundamental to the activation of this ubiquitous Ca²⁺ influx pathway.

What is the role of Orai1/CRACM1? It is a cell-membrane protein, and its constituent amino-acid sequence suggests that it has four membrane-spanning segments. Could it be the elusive CRAC channel? This is not clear yet. The amino-acid sequence does not contain the ring of negatively charged glutamate amino acids that is typical of the selectivity

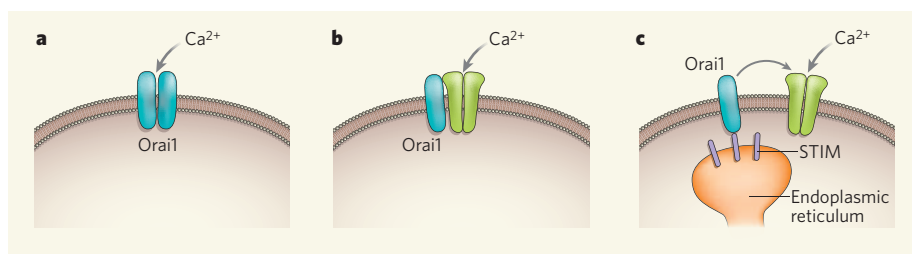


Figure 1 | Possible roles of Orai1/CRACM1 in the CRAC channel mechanism for calcium-ion influx into the cell. Feske *et al.*¹ and Vig *et al.*² have shown that the Orai1/CRACM1 protein is an essential component of the CRAC channel mechanism. **a**, Orai1/CRACM1 could be the channel itself, presumably by forming a tetramer in which four subunits come together, or **(b)** it could be a regulatory subunit of a multimeric CRAC channel complex. **c**, Alternatively, Orai1/CRACM1 could be an adaptor protein, coupling depletion of intracellular Ca^{2+} stores to opening of the CRAC channel. It may bind directly to STIM1 and then somehow open the channel, or it may sense a local signal released from the underlying endoplasmic reticulum and transduce this into channel opening.

filter of Ca^{2+} channels, nor does it show an obvious pore structure to allow passage of the ions through the membrane. However, the CRAC channel pore has unusual ion-permeation properties suggesting that it may differ from other known pores. Indeed, it is not even clear whether the CRAC complex is a channel through which ions would flow passively when it is open, or a transporter that, directly or indirectly, uses energy to move ions across the membrane. The way in which

mutations in Orai1 affect the selectivity of the CRAC channel may help resolve this issue. But even if Orai1/CRACM1 is not the entire CRAC channel, it could be an indispensable subunit of a multimeric channel complex, or it might function as a key component of the activation mechanism (Fig. 1).

There is burgeoning evidence that store-operated channels operate in non-immune cells such as endothelia, epithelia and smooth muscle. It is puzzling, therefore, that SCID

patients lacking functional CRAC channels suffer only immune disorders. An intriguing possibility is that Orai1/CRACM1 is essential for functional CRAC channels but that two closely related human proteins (Orai2 and Orai3) help to form biophysically distinct store-operated channels in these other tissues.

The shroud of mystery surrounding store-operated Ca^{2+} channels is at last being lifted by the forensic precision of molecular genetics. I anticipate rapid progress towards the complete molecular definition of CRAC channels. This will improve the prospects for developing therapeutic agents aimed at combating the growing list of human diseases associated with aberrant store-operated calcium influx⁶.

Anant B. Parekh is in the Department of Physiology, Anatomy and Genetics, University of Oxford, Parks Road, Oxford OX1 3PT, UK. e-mail: anant.parekh@physiol.ox.ac.uk

1. Feske, S. *et al.* *Nature* **441**, 179–185 (2006).
2. Vig, M. *et al.* *Science* doi:10.1126/science.1127883 (2006).
3. Putney, J. W. J. *Cell Calcium* **11**, 611–624 (1990).
4. Hoth, M. & Penner, R. *Nature* **355**, 353–356 (1992).
5. Parekh, A. B. & Penner, R. *Physiol. Rev.* **77**, 901–930 (1997).
6. Parekh, A. B. & Putney, J. W. J. *Physiol. Rev.* **85**, 757–810 (2005).
7. Roos, J. *et al.* *J. Cell Biol.* **169**, 435–445 (2005).
8. Liou, J. *et al.* *Curr. Biol.* **15**, 1235–1241 (2005).
9. Feske, S., Prakriya, M., Rao, A. & Lewis, R. S. *J. Exp. Med.* **202**, 651–662 (2005).

CHEMISTRY

A catalytic knight's move

Robert H. Crabtree

The reactivity of inert hydrocarbons can be transformed by a catalytic double act. With the ability to manipulate the lengths of the resulting carbon chains, this development opens up fresh vistas.

In chess, a knight's move can be formidable, because the implications of the combination of horizontal and diagonal motion are hard for an opponent to anticipate. Likewise, tandem reactions in chemistry are a combination of known reactions that provides an equally unexpected outcome. The power of this tandem strategy has been greatly enhanced by the use of homogeneous catalysis, in which soluble catalysts, typically metal compounds, are used to increase the rates and minimize the side-products of the individual reaction steps. In tandem catalysis^{1,2}, two or more such catalytic reactions are combined.

Writing in *Science*, Goldman, Brookhart and co-workers³ describe a novel combination of two catalytic reactions that are each celebrated reactions in their own right. The first, C–H activation, which is initiated by cleavage of a normally unreactive carbon–hydrogen bond in an organic molecule, is proving increasingly influential in organic synthesis. The second is called metathesis, and involves reactive hydrocarbons known as alkenes, which contain

carbon–carbon double bonds ($\text{C}=\text{C}$). Alkene metathesis is initiated by cleavage of such a bond, and has gained wider attention following the award of the 2005 Nobel chemistry prize to three of its pioneers, Robert Grubbs, Richard Schrock and Yves Chauvin⁴.

Both reactions have interesting histories. Initial work in the 1970s, now recognized, but at the time either ignored or even disbelieved, was carried out by Alex Shilov⁵ on C–H activation in the Soviet Union and by Chauvin⁶ on alkene metathesis in France. In each case, it has taken decades of further work by many groups to overcome the initial limitations of slow rates or lack of general applicability, until today the promise of these reactions is beginning to be more fully realized. Metathesis has moved forward farthest — in particular, the Grubbs⁷ and Schrock⁸ metathesis catalysts are widely used, because they are far more tolerant of other functional chemical groups in the reactant than were earlier catalysts.

The Goldman–Brookhart reaction³ is notable in being the first example of a combination

of these two high-profile reactions. In the C–H activation step, which is catalysed by a specially chosen iridium catalyst, normally unreactive alkanes — hydrocarbons that contain no double bonds — can be converted to their reactive alkene counterparts by losing two hydrogen atoms to form a $\text{C}=\text{C}$ double bond (Fig. 1, overleaf). A Schrock metathesis catalyst is also present, and this uses the alkene produced in the previous step as its starting point. Metathesis temporarily cleaves the strongest bond of the alkene — the $\text{C}=\text{C}$ double bond — to form fragments that are bound briefly to the catalyst. If these fragments are not identical, new products are formed when the fragments switch partners and join together, in the key phase of the metathesis process. In the third and final stage of the tandem cycle, the hydrogen released in the initial C–H activation step is added back to the $\text{C}=\text{C}$ bond of the metathesis product to give alkanes that have a different number of carbon atoms from that in the initial reactant. Because this step is simply the reverse of the first reaction in the sequence, it is catalysed by the same iridium catalyst, and no third catalyst is needed.

A surprising feature of this tandem reaction is the ability of the catalysts to work independently, without interfering with each other. This result will undoubtedly prompt other workers to try previously unconsidered catalyst combinations in the search for unexpected outcomes.

Many catalytic reactions leave the number of carbon atoms in a chain unaltered from

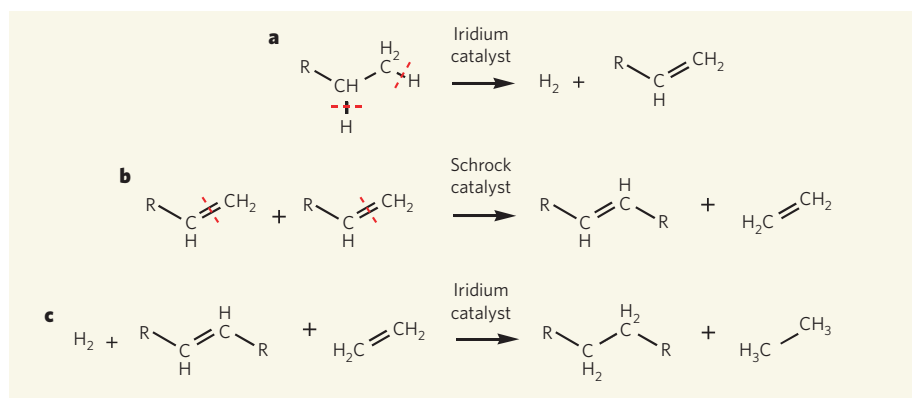


Figure 1 | Individual steps of the Goldman-Brookhart reaction³. R represents a generalized carbon chain, and the red dotted lines indicate bonds cleaved by the catalysts. **a**, A hydrocarbon loses two hydrogen atoms to form a double bond. **b**, The double bond is cleaved, and the fragments swap over in a metathesis reaction. **c**, Hydrogen from the first step adds to the metathesis products. The process gives hydrocarbons with different chain lengths compared with the starting material.

starting material to products. Metathesis is noteworthy in that it converts carbon chains into a mixture of longer and shorter chains. Thanks to the metathesis component, the same is true of the Goldman-Brookhart reaction, which involves alkanes — also known as saturated hydrocarbons — and not just compounds with C=C bonds as in traditional metathesis.

Apart from its appealing novelty, this ability

to manipulate the length of carbon chains is a potentially useful step in processing saturated hydrocarbon sources for fuel and chemical uses, particularly in a future regime aiming to use alternative feedstocks to oil, such as biomass or natural gas. The Fischer-Tropsch reaction converts mixtures of carbon monoxide and hydrogen, which may be derived from such alternative feedstocks, into a mixture of saturated hydrocarbons of various

carbon-chain lengths. Once the medium-length chains needed for gasoline are removed, longer and shorter chains remain; the Goldman-Brookhart reaction could produce desirable chain lengths from this mixture by the route shown in Figure 1. Practical application is still a goal for the future, however, because the efficiency of the alkane dehydrogenation step needs to be improved.

As oil supplies are depleted in the coming decades, energy concerns will intensify. In this new era, we will need to fully harness the great power of chemical catalysis — including tandem versions — to carry us through what might otherwise be lean times.

Robert H. Crabtree is in the Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06511, USA. robert.crabtree@yale.edu

1. Wasilke, J. C., Obrey, S. J., Baker, R. T. & Bazan, G. C. *Chem. Rev.* **105**, 1001-1020 (2005).
2. Fogg, D. E. & dos Santos, E. N. *Coord. Chem. Rev.* **248**, 2365-2379 (2004).
3. Goldman, A. S., Roy, A. H., Huang, Z., Ahuja, R., Schinski, W. & Brookhart, M. *Science* **312**, 257-261 (2006).
4. <http://nobelprize.org>
5. Shilov, A. E. & Shul'pin, G. B. *Activation and Catalytic Reactions of Saturated Hydrocarbons* (Kluwer, Dordrecht, 2000).
6. Hérisson, J. L. & Chauvin, Y. *Makromol. Chem.* **141**, 161-164 (1971).
7. Trnka, T. M. & Grubbs, R. H. *Acc. Chem. Res.* **34**, 18-29 (2001).
8. Schrock, R. R. *Acc. Chem. Res.* **23**, 158-165 (1990).

IMMUNOLOGY

What does it mean to be just 17?

Cristina M. Tato and John J. O'Shea

For a long time it was thought that there are only two types of T helper cell. But it is becoming clear that there may be other lineages that influence inflammatory responses in certain circumstances.

T cells are a workhorse of the adaptive immune system, with the job of orchestrating defences against microbial invasion. A subcategory of T cells, T helper cells, defends against microbes but also causes trouble by inducing inflammation in immune-mediated diseases. T helper cells become further specialized, or differentiated, and a type known as T_H17 has recently stepped into the limelight. These cells produce a messenger molecule called interleukin-17, and are notable both in providing defence against extracellular bacteria and in mediating inflammation.

Research on T_H17 cells is at the forefront of immunology, and the latest news comes in three papers¹⁻³. Two of them (by Mangan *et al.*¹ and Bettelli *et al.*²) appear on pages 231 and 235 of this issue, while the third (Langowski *et al.*³) has just appeared online.

Interferons and interleukins are secreted molecules, also known as cytokines, that both determine T-helper-cell fate and constitute the output of the resulting cells. Conventional

thinking was that, in combating infection, T helper cells could take two forms: T_H1 cells that produce interferon- γ (IFN- γ), and T_H2 cells that produce interleukin-4 (IL-4). The simple T_H1 - T_H2 dichotomy made sense, in that elimination of intracellular pathogens depends on IFN- γ , whereas eradication of parasitic worms requires IL-4 (Fig. 1). But autoimmune diseases do not fit into this picture, into which other subsets of T cells have made their way — first, regulatory T (T_{reg}) cells, which act to keep immune hyperactivity in check, and more recently T_H17 cells. Unexpectedly, it is becoming evident that the differentiation of T_{reg} and T_H17 cells is related.

Interleukin-17, otherwise known as IL-17A, is the founding member of a small family of cytokines, denoted IL-17A-F (ref. 4), that is generally thought to increase inflammation by recruiting other immune cells to peripheral tissues. There is now plenty of evidence that IL-17 has harmful effects in autoimmune and autoinflammatory disorders. T_H17 cells have

been proposed as a new T-cell lineage that is characterized by the selective secretion of this cytokine. But what governs T_H17 production?

Players that lie upstream in the T-helper-cell pathway are dendritic cells, and other so-called antigen-presenting cells, that provide an initial reaction to infection. This response involves the transmission of relevant cytokine signals that prompt the differentiation of specific T-cell populations. Interleukin-23, a cytokine produced by dendritic cells and other antigen-presenting cells, was thought to be required for IL-17 production and subsequent inflammatory disease⁴. Selective deficiency of IL-23 in mice attenuated immune-mediated disease, resulting in fewer IL-17-producing T cells. More recently, however, a mixture of inflammatory cytokines that includes IL-6 and transforming growth factor (TGF)- $\beta1$ was found to be a potent cocktail that results in the production of T_H17 cells⁵.

The studies described in this issue^{1,2} confirm the role of TGF- $\beta1$ and IL-6 in promoting T_H17 development, and put to rest the idea that IL-23 is the initiator of T_H17 differentiation. Specifically, TGF- $\beta1$ -deficient mice studied by Mangan *et al.*¹ had reduced numbers of T_H17 cells, although IL-17 was not completely absent in this setting. Similarly, mice studied by Bettelli *et al.*² in which TGF- $\beta1$ was overexpressed had increased numbers of T_H17 cells and more severe autoimmune disease.

This research^{1,2} provides strong evidence that TGF- $\beta1$ really is necessary for *in vivo*

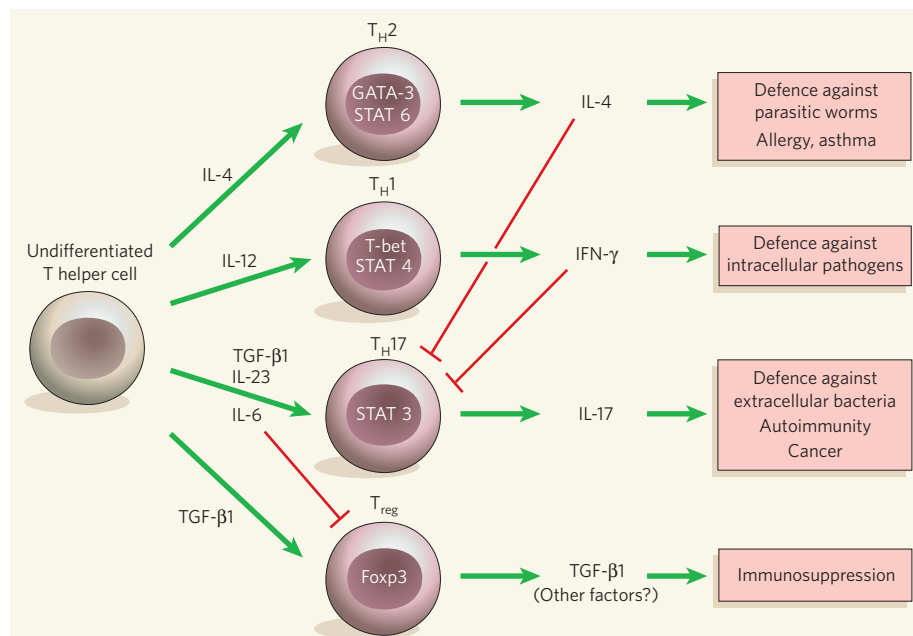


Figure 1 | T-helper-cell differentiation. Prompted by different types of interleukin (IL) produced by dendritic cells and other sources, undifferentiated T helper cells can develop into the T_H1 or T_H2 lineages. In an inflammatory response, TGF- β 1 and IL-6 promote the development of another lineage, T_H17 cells that produce IL-17. In contrast, interferon- γ (IFN- γ) and IL-4, products of T_H1 and T_H2 cells, inhibit T_H17 differentiation. TGF- β 1 boosts expression of the IL-23 receptor, promoting expansion of T_H17 cells by IL-23. But TGF- β 1 also promotes the development of another lineage — regulatory T (T_{reg}) cells — by inducing the transcription factor Foxp3, an outcome that is inhibited in the presence of IL-6. Development of T_H1 and T_H2 cells depends on specific STAT proteins and other gene-transcription factors such as T-bet and GATA-3. STAT-3 is probably involved in T_H17 differentiation, but other T_H17 -lineage-specific factors may well emerge.

T_H17 differentiation. However, whether other cytokines are required is unresolved. In these studies, T_H17 cells were most effectively produced in mixed cell cultures where T cells were stimulated in the presence of antigen-presenting cells. Factors other than IL-6 and TGF- β 1 may also be important for this lineage. Moreover, many cells produce TGF- β 1 and IL-6, and the main sources of these cytokines for T_H17 differentiation remain unclear.

What about IL-23? The receptor for IL-23 is not expressed on precursor T cells, but Mangan *et al.*¹ report that its expression is induced by TGF- β 1. So although IL-23 is not an instigator of T_H17 differentiation, it may be involved in expanding and maintaining this population of cells.

In all, the new studies^{1,2} point to the existence of another dichotomy in T-cell differentiation. By itself, TGF- β 1 induces the differentiation of T_{reg} cells. The combination of IL-6 in conjunction with TGF- β 1 favours IL-17-producing cells, but inhibits the expression of Foxp3, a gene-transcription factor that is essential for T_{reg} differentiation. Thus, T_H17 and T_{reg} cells are evidently related, and represent two sides of an inflammatory coin regulated by the same cytokine. The opposing effects of TGF- β 1 may seem confusing, but other cytokines have complex dual actions⁶. It is of interest that there were fewer T_{reg} cells in autoimmune disease associated with TGF- β 1 overexpression; perhaps IL-17 antagonizes Foxp3 expression. The counter-regulation of

T_{reg} and T_H17 will undoubtedly be further analysed.

If IL-23 and IL-17 can be such troublemakers, are they beneficial in host defence? Yes, is the answer. Interleukin-17E has a pivotal role in protection against parasitic worms^{7,8}; and Mangan *et al.*¹ show that IL-23 is necessary for resistance to *Citrobacter rodentium*, an extracellular bacterial pathogen related to pathogenic *Escherichia coli*, and the authors argue that IL-23 is important for an effective IL-17 response *in vivo*. Blocking TGF- β 1 also exacerbated disease, again consistent with the importance of this cytokine in *in vivo* regulation of IL-17. Infection with *C. rodentium* was associated with some cells that produced both IL-17 and IFN- γ — a notable observation, because the production of these two cytokines was thought to be mutually exclusive.

Clearly, the relationships between T_H1 , T_{reg} and T_H17 cells will need to be further assessed, and much attention has already centred on the gene-transcription factors that drive T-cell differentiation (Fig. 1). Differentiation of T_H17 cells seems to be independent of STAT-4 and STAT-6 (transcription factors that regulate T_H1 and T_H2 cells, respectively). STAT-3 is activated by both IL-23 and IL-6, and binds to the IL-17 and IL-17F gene promoters⁹. But it remains to be seen whether T_H17 cells selectively express lineage-specific transcription factors analogous to two others (T-bet and GATA-3) that are intimately involved in T_H1 and T_H2 differentiation.



50 YEARS AGO

The longest earthworm in the world, *Megascolides australis*, is found in Gippsland, Australia, and grows up to eleven feet in length... When disturbed, the worm squirts out a series of pairs of jets of fluid from a line of pores opening down each side of the body. The effect can be most spectacular, for these jets rise as high as eighteen inches or two feet into the air. Although there are reports that the fluid has a corrosive action, it is only slightly alkaline and contains some dissolved salts, body wastes like urea and some proteinous materials and cells. The fluid comes from the worm's body cavity and is squirted out by violent contractions of the body-wall which force the fluid out under great pressure through the pores. There is no record of the fluid having anything but a mildly irritating effect on the skin of human beings. The fluid is used for lining or lubricating the burrows of the worms.

From *Nature* 12 May 1956.

100 YEARS AGO

"The bicentenary celebration of the birth of Benjamin Franklin" — The oldest scientific society in the new world is, I believe, the American Philosophical Society of Philadelphia...founded by Benjamin Franklin... An altogether exceptional feature of the ceremony was that a degree was conferred on the King, who was represented by Sir Mortimer Durand, H.M. Ambassador at Washington. In announcing this degree the Provost read with great effect the celebrated speech on England from Henry V. It is pleasant to record the enthusiastic cheers which the whole audience gave, standing, as the Ambassador was hooded... An American dinner is managed somewhat differently from our own, for the toast-master is not, as with us, a servant with a stentorian voice, but is the most highly honoured of the hosts of the occasion... Those who have taken part in such festivals in America need not be told that the organization was admirable and the hospitality unbounded.

From *Nature* 10 May 1906.

50 & 100 YEARS AGO

Finally, understanding the regulation of IL-17 takes on additional significance because of a newly identified connection between inflammation and cancer³. The paper concerned, by Langowski *et al.*³, implicates IL-23 and IL-17 in causing cancer: the authors report that IL-23 and IL-17 are increased in human tumours, and that IL-23 deficiency is associated with fewer tumours in mice that are liable to develop cancer.

It must seem to those outside the business that research developments in immunology merely add complications to our understanding of the immune system. The new papers^{1–3} do indeed add another layer of complexity to T-cell differentiation, and leave many open questions. But they also present exciting data that connect IL-17 regulation with human diseases that range across autoimmunity,

infection and cancer, and as such they offer new avenues for the development of treatments. ■

Cristina M. Tato and John J. O'Shea are in the Molecular Immunology and Inflammation Branch, National Institute of Arthritis, Musculoskeletal and Skin Diseases, Bethesda, Maryland 20892-1820, USA.
e-mails: tatoc@mail.nih.gov;
sheaj@arb.niams.nih.gov

1. Mangan, P. R. *et al.* *Nature* **441**, 231–234 (2006).
2. Bettelli, E. *et al.* *Nature* **441**, 235–238 (2006).
3. Langowski, J. L. *et al.* doi:10.1038/nature04808 (2006).
4. Dong, C. *Nature Rev. Immunol.* **6**, 329–333 (2006).
5. Veldhoen, M., Hocking, R. J., Atkins, C. J., Locksley, R. M. & Stockinger, B. *Immunity* **24**, 179–189 (2006).
6. O'Shea, J. J., Ma, A. & Lipsky, P. *Nature Rev. Immunol.* **2**, 37–45 (2002).
7. Fallon, P. G. *et al.* *J. Exp. Med.* **203**, 1105–1116 (2006).
8. Owyang, A. M. *et al.* *J. Exp. Med.* **203**, 843–849 (2006).
9. Chen, Z. *et al.* *Proc. Natl Acad. Sci. USA* (in the press).

THERMODYNAMICS

When a phase is born

Pablo G. Debenedetti

Phase changes in matter generally occur by building up from small nuclei of the new phase. Scattering experiments and computer simulations reveal the characteristic size of the smallest of these nuclei.

Rain and snowfall, petroleum refining and the purification of pharmaceuticals by crystallization are examples of natural and industrial processes that involve a phase transition. Understanding the molecular mechanisms underlying these ubiquitous events is therefore of widespread interest. Writing in *Journal of Physical Chemistry B*, Albert C. Pan and colleagues¹ suggest a novel way of measuring the characteristic size of the smallest building-block of a new phase that can grow spontaneously, and show how this quantity changes across thermodynamic conditions ranging from metastability to instability.

In phase transitions such as crystallization and boiling, the two phases involved have different densities (ice floats on water) and entropies (heat must be added to water to make it boil). Such transitions are known as first-order transitions, and they require that the starting, or mother phase, have a higher free energy (energy available to do work) than the new phase at the same temperature and pressure. In this situation, the mother phase — for example, liquid water below freezing — will give way to the new phase, ice. The mother phase is said to be metastable with respect to the new phase, and the degree of departure of the system from equilibrium conditions is referred to as its supersaturation.

First-order phase transitions occur through a nucleation mechanism², in which a molecular aggregate of the new phase begins to form within the mother phase. The lower free

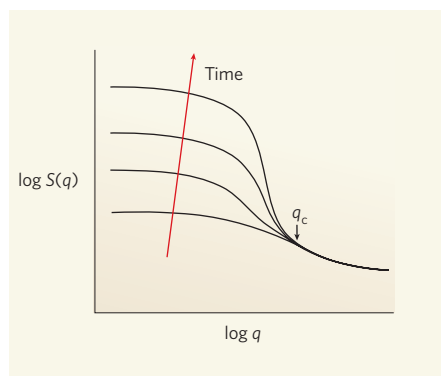


Figure 1 | The search for the critical nucleus. Scattering intensity, or structure factor $S(q)$, as measured by Pan and colleagues¹ for nucleation in a phase-separating mixture. Curves obtained at different times during a phase transition merge at a critical scattering vector q_c , implying that the length $1/q_c$ is a signature of the critical nucleus from which the new phase forms.

energy of the new phase favours nucleus formation, but the process is hindered by the energetic cost of forming an interface between the nucleus and the mother phase. The driving force for nucleus formation is proportional to the volume of the nucleus, whereas the cost of forming an interface is proportional to its surface area. The former will thus balance the latter only for a sufficiently large nucleus, and nuclei larger than a critical size grow spontaneously, whereas those smaller than this threshold dissolve back into the mother phase.

This picture of a nucleus as a localized and rare fluctuation underlies classical nucleation theory², and is an adequate model at small supersaturations. Under these conditions, the nucleation time is long relative to the structural relaxation time, and the nucleation process can be treated by equilibrium (thermodynamic) considerations involving the work needed to form the critical nucleus.

Pan *et al.*¹ use structure factors that they obtain in neutron-scattering experiments on phase-separating polyolefin blends to determine the length scale associated with the critical nucleus, a procedure developed originally by co-author Nitash Balsara and his group^{3,4}. The structure factor, $S(q)$, is the intensity distribution, as a function of the scattering vector, q , of probing neutrons that have been deflected off molecules in the mixture⁵. This scattering vector is given by $q = (4\pi/\lambda)\sin(\theta/2)$, where θ is the scattering angle and λ the wavelength of the incident neutrons. The growth of structures with a characteristic size l is associated with an increase in scattering, and so the structure factor, at $q \sim 1/l$.

Structure-factor curves obtained at different times during a nucleation process merge at a critical scattering vector q_c (Fig. 1). In other words, for scattering vectors larger than q_c , the scattering intensity is independent of time, and structures of characteristic size smaller than $1/q_c$ do not grow in time. The merging of curves at q_c is thus a signature of the size of the critical nucleus^{3,4}. The fact that the characteristic size of critical nuclei can be determined from the structure factor is surprising: it suggests that the fluctuations that give rise to these nuclei resemble ordinary concentration fluctuations more than the localized, high-intensity fluctuations envisioned in classical nucleation theory.

The authors also calculate the structure factor through computer simulation of a model phase-separating binary mixture. The measured and computed structure factors are qualitatively very similar. Furthermore, the critical nucleus size obtained from the computed structure factor agrees with that obtained independently by calculating the free-energy cost of forming nuclei of different sizes, the maximum of such a free-energy curve corresponding to the critical nucleus. The results provide a direct test of the relationship between the critical nucleus size and the intensity distribution of scattered radiation, and so point to a new way of measuring the size of a critical nucleus. As this quantity is difficult to obtain experimentally, this is an important development. The structure factor by itself, however, does not yield direct information on the number of molecules in the critical nucleus.

At large supersaturations, the interface between nucleus and mother phase becomes progressively diffuse. In this situation, theory focuses on the free-energy cost of establishing a non-sharp interface^{6,7}. According to the classical theory, the size of the critical nucleus

decreases steadily with increasing supersaturation, produced using different values of temperature or pressure, and remains finite at the 'spinodal' point where the mother phase becomes thermodynamically unstable. In contrast to phase transitions that originate from a metastable phase, transitions from an unstable mother phase ('spinodal decomposition') occur spontaneously, bypassing the need for critical nuclei to form. Theories that approximate interfacial thermodynamics by neglecting density or concentration fluctuations predict a sharp transition from nucleation to spinodal decomposition and a diverging critical nucleus at the spinodal⁶. Scaling arguments, in contrast, predict a smooth transition from metastability to instability⁸. In agreement with recent experiments on concentrated protein solutions⁹, Pan *et al.*¹ find that the size of the critical nucleus decreases smoothly with supersaturation and remains finite at the spinodal. They see no evidence of a diverging critical nucleus size.

An important open question is the geometry of the critical nucleus. Computer simulations¹⁰ indicate that, at large supersaturations, nuclei send out offshoots to become ramified, fractal objects. But the definition of the critical nucleus size used by Pan *et al.* in their simulations assumes a compact object, and it would be interesting to account for the shape of the critical nucleus in future calculations.

Although a qualitative picture of the transition from nucleation to spinodal decomposition exists⁷, this has yet to result in a predictive theory of nucleation at large supersaturations. Because of the ambiguity associated with the very notion of a critical nucleus under these conditions⁷, such a theory will probably require kinetic arguments (based on the rates of growth and decay of embryonic nuclei)¹¹, rather than thermodynamic arguments based on the free-energy cost of forming a critical nucleus. The work of Pan *et al.* provides microscopic insight on the smallest building-block of a new phase, an important ingredient for future theoretical descriptions. ■

Pablo G. Debenedetti is in the Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544, USA.
e-mail: pdebene@princeton.edu

1. Pan, A. C., Rappl, T. J., Chandler, D. & Balsara, N. P. *J. Phys. Chem. B* **110**, 3692–3696 (2006).
2. Debenedetti, P. G. *Metastable Liquids: Concepts and Principles* (Princeton Univ. Press, 1996).
3. Lefebvre, A. A., Lee, J. H., Balsara, N. P. & Hammouda, B. *J. Chem. Phys.* **116**, 4777–4781 (2002).
4. Lefebvre, A. A., Lee, J. H., Balsara, N. P. & Vaidyanathan, C. *J. Chem. Phys.* **117**, 9063–9073 (2002).
5. Stanley, H. E. *Introduction to Phase Transitions and Critical Phenomena* (Oxford Univ. Press, 1971).
6. Cahn, J. W. & Hilliard, J. E. *J. Chem. Phys.* **31**, 688–699 (1959).
7. Wood, S. M. & Wang, Z.-G. *J. Chem. Phys.* **116**, 2289–2300 (2002).
8. Binder, K. *Physica* **140A**, 35–43 (1986).
9. Shah, M., Galkin, O. & Vekilov, P. G. *J. Chem. Phys.* **121**, 7505–7512 (2004).
10. Heerman, D. W. & Klein, W. *Phys. Rev. Lett.* **50**, 1062–1065 (1983).
11. Ruckenstein, E. & Dijkstra, Y. S. *J. Coll. Interf. Sci.* **118**, 51–72 (2005).

ANALYTICAL CHEMISTRY

Cause for a llama

The economic value of the even-toed ungulates of the camelid family known as llamas (pictured) has been principally founded on their use as pack animals and as a source of hide and wool in their native South America. That llama antibodies are suited to measuring levels of caffeine in drinks had been overlooked, an omission now corrected by Ruth C. Ladenson and colleagues (*Anal. Chem.* doi:10.1021/ac058044j; 2006).

Although caffeine-specific antibodies for use in immunoassays are available commercially, they become irreversibly denatured at high temperatures. But camelids produce certain antibodies that consist exclusively of heat-resistant chains of amino acids. Such antibodies would be ideal for assessing the caffeine content of both hot and cold beverages.



F. GOHIER/ARDEA.COM

Over a period of ten weeks, the authors immunized five camelids — three llamas and two camels — with caffeine linked to an immune stimulant called keyhole limpet haemocyanin. They subsequently identified and isolated antibody fragments that bind only to caffeine in blood samples from the immunized animals, and produced one of these as a soluble protein.

The chosen antibody preparation retained almost all of its activity after incubation at

temperatures up to 90 °C, whereas that of comparable preparations derived from mice dropped away sharply above 70 °C. Swilled through regular and decaffeinated coffee and decaffeinated soft drinks, in each case it registered caffeine levels in good agreement with established values.

The authors' next step is to produce a dipstick assay, similar to that used in home pregnancy tests, for point-of-consumption use.

Richard Webb

GEOCHEMISTRY

The noble art of recycling

Takuya Matsumoto

Xenon trapped beneath Earth's crust provides clues to how our planet evolved, but quantifying atmospheric contamination has been impossible. The latest analysis surmounts a barrier to our understanding.

Because of their scarcity, chemical stability and presence as many distinguishable isotopes, noble gases in Earth's mantle — the solid layer between its outer crust and liquid core — provide constraints on the origin, structure and evolution of Earth and its atmosphere. On page 186 of this issue¹, Holland and Ballentine use investigations of well gases from the upper mantle to challenge an established tenet concerning noble-gas abundance: the existence of a 'subduction barrier' that prevents the noble gases from recycling into the mantle through tectonic activity. In fact, the authors conclude that some 80% of xenon in the mantle comes from sea water introduced by just such a process.

The subduction barrier was described in a classic paper² showing that at subduction zones — where one of Earth's tectonic plates dives under another — the noble gases in subducting

materials are returned to the surface through volcanic activity, leaving the composition of the mantle unaffected. The fact that noble gases of the same isotopic composition as Earth's atmosphere are found in nearly all analyses of samples of rock extruded from the mantle has been attributed, quite reasonably, to contamination from the atmosphere during, or after, the eruption of magma^{3,4}. With the exception of certain mantle-derived rocks that acquire fluid at subduction-zone settings^{5,6}, the recycling of noble gases into the mantle reservoirs by subduction has never been proved.

Holland and Ballentine¹ analyse high-quality isotope data from well gases in New Mexico to demonstrate that xenon of atmospheric nature is intrinsic to the mantle. They study the abundances of three primordial isotopes of xenon — ¹²⁴Xe, ¹²⁶Xe, ¹²⁸Xe, which are not produced in any reaction inside Earth —

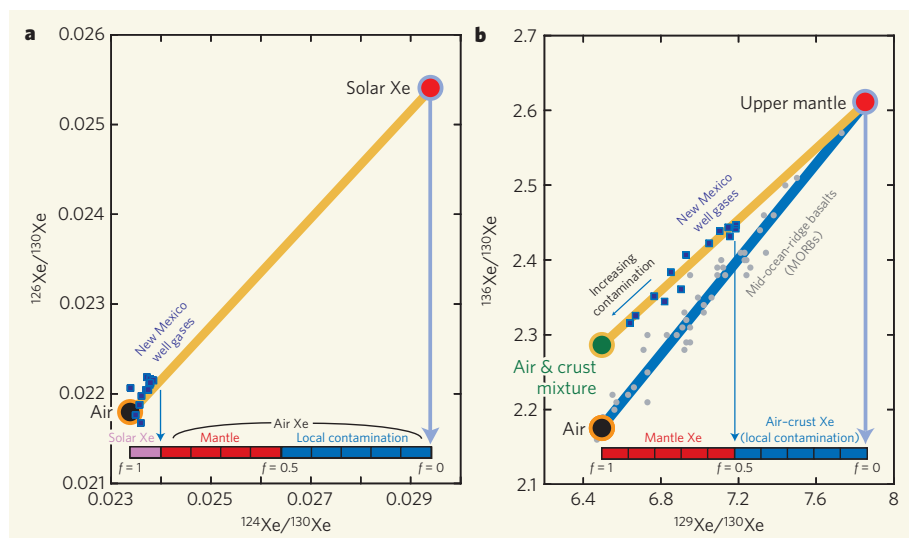


Figure 1 | Differing isotopic ratios of xenon. **a**, The ratios $^{124}\text{Xe}/^{130}\text{Xe}$ and $^{126}\text{Xe}/^{130}\text{Xe}$ in the Sun and in Earth's atmosphere are well known (large circles). Holland and Ballentine's data¹ from gas wells in New Mexico (blue squares) fit a straight line drawn between these two points, indicating samples with variable components from solar- and air-like sources. The fraction, f , of air ^{130}Xe in each sample can be estimated from these data by solving a mixing equation (scale bar). Part of the required air-like xenon (red on the scale bar) cannot be accounted for by local air contamination. **b**, The ratios $^{129}\text{Xe}/^{130}\text{Xe}$ and $^{136}\text{Xe}/^{130}\text{Xe}$ of New Mexico well gas (blue squares) and MORB rock (grey circles)^{9,12} plot on different mixing lines: whereas MORB data extend from air composition, suggesting that they contain purely atmospheric contamination, data from New Mexico well gases extend from a point consistent with pre-mixed air–crust gases, which are enriched in ^{136}Xe from crustal uranium fission. The intersection of the two lines defines the isotopic ratios for the mutual source of the two samples — most probably the convecting upper mantle.

in relation to that of a fourth primordial isotope, ^{130}Xe , and compare the ratios obtained with those found in xenon in the Sun and in Earth's atmosphere (Fig. 1a). Their results are consistent with a xenon content of 90% atmospheric and 10% solar origin.

The authors address the origin of the air-like xenon component using the ratios of two further isotopes, ^{129}Xe and ^{136}Xe , to ^{130}Xe . These are radiogenic isotopes, respectively produced by radioactive decay from a now-extinct iodine isotope (^{129}I), and by fission of a now-extinct plutonium isotope (^{244}Pu) and long-lived uranium-238. Non-atmospheric $^{129}\text{Xe}/^{130}\text{Xe}$ and $^{136}\text{Xe}/^{130}\text{Xe}$ ratios have already been established in data from extruded mantle rocks known as mid-ocean-ridge basalts (MORBs). These higher ratios can probably be explained by assuming that significant amounts of primordial ^{130}Xe were removed from the mantle source by early catastrophic degassing. Quantifying the 'pristine' mantle composition will therefore constrain models of Earth's evolution from its early stages to the development of distinct mantle reservoirs⁷. But the pristine composition cannot be identified from MORB data alone, as the degree of atmospheric contamination after the rocks had been extruded is unknown.

The authors¹ find that, on a three-isotope diagram, $^{129}\text{Xe}/^{130}\text{Xe}$ and $^{136}\text{Xe}/^{130}\text{Xe}$ ratios of the New Mexico well gases form a straight line that does not extend from the point that represents those ratios in air (Fig. 1b). Instead, the data lie on a line that extends from a point

consistent with a hybrid of atmospheric xenon and xenon of crustal origin. The point of intersection of this line with that of the MORB data will represent the xenon composition of the common source of MORBs and well gases, most probably the upper mantle.

The results from the radiogenic isotopes indicate that the air–crust hybrid component, which is likely to be added to well gases through local contamination during their storage in the continental crust^{1,8}, can account for only about half of the 90% atmospheric xenon component revealed by the non-radiogenic xenon isotopes. The rest of this air xenon must be intrinsic to the mantle reservoir, and the most likely process to introduce it into the mantle source region would be subduction.

Thus it seems that the noble-gas subduction barrier is not as effective as had been presumed. The survival of only 2% of pore water in subducting oceanic plates would be sufficient to account for the quantity of isotopically air-like noble gases found in the upper mantle¹. But can these gases be recycled to the deeper mantle regions? Rocks known as oceanic island basalts (OIBs), which are presumed to be of lower-mantle origin, yield noble-gas isotopic ratios that are less radiogenic (and thus more air-like) than MORBs. The difference is normally explained by assuming that MORBs originate from a degassed reservoir, from which the primordial gas has largely been removed to form the present atmosphere.

Holland and Ballentine speculate, however,

that the difference in noble-gas isotope signatures between the MORBs and OIBs might in fact result from different degrees of regassing by subduction. Such a model requires that noble gases in the OIB source should be more significantly modified by the recycled component than are those in the MORB source, but this is quite consistent with results from other geochemical tracers such as neodymium, strontium and lead isotopes. Recent attempts to determine precisely the ratios of primordial xenon isotopes in OIBs and MORBs could also support the idea that subduction adds an intrinsic air component to the source regions of these rocks^{9–11}.

Holland and Ballentine explore the details of this air influx by showing that the recycled air found in the New Mexico well gases has relative noble-gas abundance ratios similar to those of sea water. But although sea water is undoubtedly a ready source of air-like noble gases in subducting materials, it is unclear how the seawater signature could survive subduction, storage in the mantle or crustal reservoirs and emanation at the continental gas field, without changes to its relative noble-gas abundances. Even the least-contaminated MORBs do not⁹ contain solar-like xenon, so the underlying assumption that MORBs and the New Mexico samples share the same uniform mantle reservoir needs further clarification. There are also many unresolved issues about the feasibility of deep air recycling — for example, the presence of a carrier in subducting materials other than pore water that can relay air-like noble gas into the deeper mantle regions.

Nevertheless, supported by an independent determination of xenon isotopic composition in the upper mantle¹², Holland and Ballentine's conclusion¹ that some air-like xenon in the New Mexico well-gas samples is intrinsic to their mantle source seems robust. However, extending the recycling hypothesis to other samples, such as MORBs and OIBs, with their uncertain air contamination, will require further work.

Takuya Matsumoto is in the Department of Earth and Space Science, Graduate School of Science, Osaka University, Toyonaka 560-0043, Japan.
e-mail: matsumoto@ess.sci.osaka-u.ac.jp

- Holland, G. & Ballentine, C. J. *Nature* **441**, 186–191 (2006).
- Staudacher, T. & Allègre, C. J. *Earth Planet. Sci. Lett.* **89**, 173–183 (1988).
- Patterson, D. B., Honda, M. & McDougall, I. *Geophys. Res. Lett.* **17**, 705–708 (1990).
- Ballentine, C. J. & Barford, D. N. *Earth Planet. Sci. Lett.* **180**, 39–48 (2000).
- Matsumoto, T., Chen, Y. & Matsuda, J. *Earth Planet. Sci. Lett.* **185**, 35–47 (2001).
- Matsumoto, T. et al. *Earth Planet. Sci. Lett.* **238**, 130–145 (2005).
- Trieloff, M. & Kunz, J. *Phys. Earth Planet. Inter.* **148**, 13–38 (2005).
- Ballentine, C. J., Sherwood Lollar, B., Marty, B. & Cassidy, M. *Nature* **433**, 33–38 (2005).
- Kunz, J., Staudacher, T. & Allègre, C. J. *Science* **280**, 877–880 (1998).
- Trieloff, M., Kunz, J. & Allègre, C. J. *Earth Planet. Sci. Lett.* **200**, 297–313 (2002).
- Trieloff, M. et al. *Science* **288**, 1036–1038 (2000).
- Moreira, M., Kunz, J. & Allègre, C. J. *Science* **279**, 1178–1181 (1998).

BRIEF COMMUNICATIONS

Diving insects boost their buoyancy bubbles

Underwater backswimmers use their haemoglobin to help them stay stationary while waiting for prey.

Backswimmers (Notonectidae) are common diving insects found around the world that exploit the mid-water zone for predation — they breathe by using an air bubble collected at the surface. Here we show that backswimmers achieve prolonged periods of neutral buoyancy by using oxygen stored in their haemoglobin to stabilize the volume of the bubble as they breathe from it. This enables them to maintain their position in the water column without continually swimming.

Many air-breathing aquatic insects dive with air bubbles that either supply oxygen directly or function as a 'gas gill' that obtains dissolved oxygen from the water^{1,2}. But bubble volume decreases during a dive because of oxygen uptake by the insect and diffusion of carbon dioxide and nitrogen into the surrounding water. Most diving insects are therefore limited to floating at the water's surface or clinging to submerged bodies.

Backswimmers are remarkable in that they are the only insects to inhabit mid-water environments as adults and are also the only ones to have haemoglobin throughout their entire life cycle³. Although it has been suggested that haemoglobin must play a role in buoyancy control^{4–6}, this has not been directly confirmed. By using new technology involving a sensitive electronic balance and fibre-optic



Hold your breath: the backswimmer *Anisops deanei* uses an air bubble to survive underwater.

oxygen sensors (for details of methods, see supplementary information), we have been able to measure changes in buoyancy and oxygen partial pressure (p_{O_2}) within the bubble of the Australian backswimmer *Anisops deanei*.

After the bubble is captured, buoyancy and p_{O_2} decrease linearly; these then plateau until the insect starts to swim to the surface to renew the bubble (Fig. 1). The insect obtains oxygen mainly from the bubble during the

initial linear phase, but from its haemoglobin during the plateau phase. The plateau begins at a p_{O_2} of 4.3 kilopascals (kPa) and ends at 3.3 kPa (5 replicates), which corresponds to a haemoglobin–oxygen affinity (P_{50}) of 3.9 kPa — similar to that measured in *A. assimilis*⁶. Adding 15% carbon monoxide to the breathing gas prevents the insect's haemoglobin from combining with oxygen and eliminates the plateau (Fig. 1), indicating that oxygen released from haemoglobin stabilizes p_{O_2} within the bubble for about 4 min. Note that backswimmers do not use their bubbles as gas gills⁶, so nitrogen loss from the bubble is minimal.

Neutral buoyancy is achieved only when the buoyant force of the bubble balances the submerged weight of the insect. Neutral buoyancy is not attained immediately upon diving: we found that when backswimmers begin a dive, they carry a bubble that is 17% larger (s.e. = 3%, $n = 10$) than would be required for neutral buoyancy. This overinflation is related to the properties of *A. deanei* haemoglobin: its high oxygen affinity prevents it from releasing its oxygen until the p_{O_2} has dropped from 20.6 kPa to about 4.3 kPa, which is equivalent to a 16% reduction in the initial volume of the bubble. The initial volume therefore provides the maximum amount of oxygen in the bubble at the start of the plateau period. Unrestrained backswimmers must swim against positive buoyancy initially, and this activity causes a steeper decrease in p_{O_2} than in our experiments, where the insects' metabolic rates were lower because they were motionless.

Philip G. D. Matthews, Roger S. Seymour

Environmental Biology, University of Adelaide, Adelaide 5005, Australia

e-mail: philip.matthews@adelaide.edu.au

1. Ege, R. Z. *Allg. Physiol.* **17**, 81–125 (1915).
2. Thorpe, W. H. & Crisp, D. J. *J. Exp. Biol.* **24**, 227–269 (1947).
3. Miller, P. L. *J. Exp. Biol.* **44**, 529–543 (1966).
4. Bare, C. O. *Univ. Kansas Sci. Bull.* **18**, 265–349 (1929).
5. Miller, P. L. *Nature* **201**, 1052 (1964).
6. Wells, R. M. G., Hudson, M. J. & Brittain, T. J. *Comp. Physiol.* **142**, 515–522 (1981).

Supplementary information accompanies this communication on Nature's website.

Received 23 January; accepted 27 March 2006.

Competing financial interests: declared none.

doi:10.1038/441171a

BRIEF COMMUNICATIONS ARISING online
www.nature.com/bca see Nature contents.

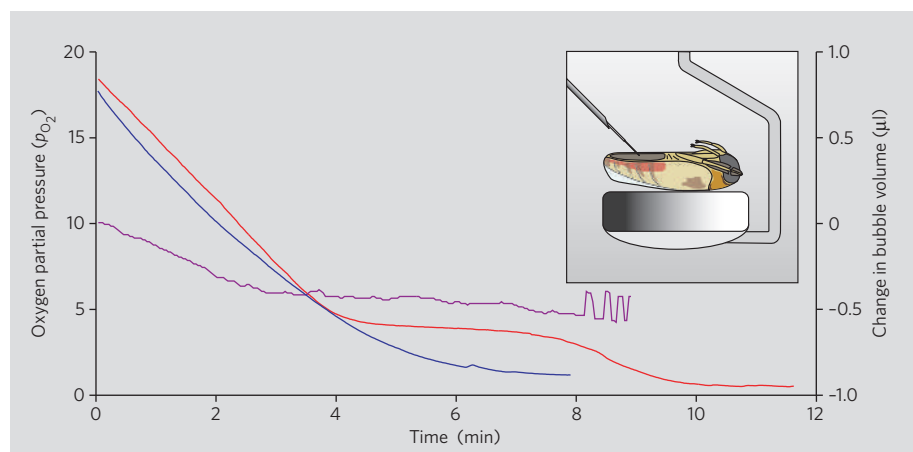


Figure 1 | Partial pressure of oxygen inside a bubble and the change in the bubble's volume while attached to the insect *Anisops deanei* during a dive. Oxygen released from haemoglobin stabilizes bubble p_{O_2} (red line) and volume (purple line) after 4 min in the motionless insect. Fluctuations in buoyancy (purple line) at 8 min indicate attempts to swim to the surface. Exposure to 15% carbon monoxide before a dive prevents the insect's haemoglobin from binding and releasing oxygen (blue line), eliminating the p_{O_2} and buoyancy plateaux. Inset, experimental setup showing *A. deanei* secured on a submerged platform attached to a balance; an optical setup probe is inserted into the abdominal air bubble.

CORRIGENDUM**Sporting contests: Seeing red? Putting sportswear into context**

C. Rowe, J. M. Harris, S. C. Roberts

Nature **437**, doi:10.1038/nature04306 (2005)

It has been drawn to our attention by David Matsumoto and Stephanie Hata that our first-round analysis of the Athens 2004 judo competition was biased by non-random allocation of blue *judogis* to a small number of seeded competitors. Although only 25 of 301 contests analysed are affected, the assumptions upon which those analyses were based were therefore incorrect. However, their reanalysis (D. M. and S. H., unpublished results), which excludes first-round matches, supports our conclusion that winning biases exist for male judo competitors wearing blue. Our interpretation of visibility-influenced winning biases therefore remains a valid and plausible alternative to that of Hill and Barton (R. A. Hill and R. A. Barton *Nature* **435**, 293; 2005).

doi:10.1038/nature04848

Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4

Mark Gilchrist¹, Vesteynn Thorsson¹, Bin Li¹, Alistair G. Rust¹, Martin Korb¹, Kathleen Kennedy¹, Tsonwin Hai², Hamid Bolouri¹ & Alan Aderem¹

The innate immune system is absolutely required for host defence, but, uncontrolled, it leads to inflammatory disease. This control is mediated, in part, by cytokines that are secreted by macrophages. Immune regulation is extraordinarily complex, and can be best investigated with systems approaches (that is, using computational tools to predict regulatory networks arising from global, high-throughput data sets). Here we use cluster analysis of a comprehensive set of transcriptomic data derived from Toll-like receptor (TLR)-activated macrophages to identify a prominent group of genes that appear to be regulated by activating transcription factor 3 (ATF3), a member of the CREB/ATF family of transcription factors. Network analysis predicted that ATF3 is part of a transcriptional complex that also contains members of the nuclear factor (NF)- κ B family of transcription factors. Promoter analysis of the putative ATF3-regulated gene cluster demonstrated an over-representation of closely apposed ATF3 and NF- κ B binding sites, which was verified by chromatin immunoprecipitation and hybridization to a DNA microarray. This cluster included important cytokines such as interleukin (IL)-6 and IL-12b. ATF3 and Rel (a component of NF- κ B) were shown to bind to the regulatory regions of these genes upon macrophage activation. A kinetic model of *Il6* and *Il12b* messenger RNA expression as a function of ATF3 and NF- κ B promoter binding predicted that ATF3 is a negative regulator of *Il6* and *Il12b* transcription, and this hypothesis was validated using *Atf3*-null mice. ATF3 seems to inhibit *Il6* and *Il12b* transcription by altering chromatin structure, thereby restricting access to transcription factors. Because ATF3 is itself induced by lipopolysaccharide, it seems to regulate TLR-stimulated inflammatory responses as part of a negative-feedback loop.

The innate immune system is the body's first line of defence against infection. It identifies foreign invaders using pattern recognition receptors, which detect highly conserved microbial-specific structures (PAMPs)^{1–4}; the TLRs are prototypic pattern recognition receptors⁵. Macrophages, activated via TLRs, unfold a tightly controlled pathogen-specific immune response⁶. Much is known about the activation of macrophages leading to the transcription of single genes. For example, it is well known that the transcription factor NF- κ B has a central role in TLR4-induced transcription of cytokines such as IL-6 and IL-12b. However, TLR activation involves a complex transcriptional programme with changes in >1,000 genes, and our overall understanding of both positive and negative transcriptional control is woefully inadequate⁷. The tools of systems biology are well suited to investigate the complex interactions induced in macrophages by TLR activation; global transcription can be measured using complementary DNA microarrays, and computational analysis of this information can lead to a deeper understanding of the system as a whole^{2,3}. We report here on a novel self-regulatory mechanism in the TLR pathway identified by global analysis of successive waves of transcriptional activity induced during macrophage activation.

Transcriptional profiling of LPS-stimulated macrophages

Temporal activation of macrophages by the TLR4 agonist bacterial lipopolysaccharide (LPS) was analysed using cDNA microarrays. These data were clustered to reveal prominent groups of genes with

similar changes in expression pattern using a *k*-means algorithm (Fig. 1). Eleven clusters comprising regulated 'waves' of transcription with early, intermediate and late phases were defined. We were particularly interested in identifying early clusters of transcription factors, because these were likely to control subsequent rounds of transcriptional activation. Cluster 6 met these criteria; it contained several transcription factors (ATF3, Btg2, Egr1, Egr2, Fos, Ier2, Jun and Rel) whose mRNA expression peaked at 1 h (Fig. 1 and Supplementary Data 1). We hypothesized that these clustered genes are co-regulated⁸ and that they share *cis*-regulatory elements. *Cis*-regulatory analysis using MotifMogul (<http://labs.systemsbio.net/bolouri/Mogul/>; see below) predicted over-representation of ATF/CREB binding sites in cluster 6. Although it is not yet possible to differentiate computationally between the DNA binding sites of the ATF/CREB transcription factors, ATF3 was the likely candidate as it was the only member of the family that was induced after LPS stimulation of macrophages (Supplementary Data 2, Fig. 2S1). ATF3 has not previously been implicated in regulation of the immune response. It is a member of the CREB family of basic leucine zipper transcription factors, and has been shown to act both as a transcriptional activator or repressor depending on the cell type and stimulus⁹. The biological role of ATF3 is also obscure. It has been reported to function in the stress response, the regulation of the cell cycle, and in apoptosis^{9–14}, although the details of how this occurs are scant and contradictory. For example, ATF3 has been implicated both as a tumour suppressor¹⁰ and as an

¹Institute for Systems Biology, Seattle, Washington 98103, USA. ²Department of Molecular and Cellular Biochemistry, Ohio State University, Columbus, Ohio 43210, USA.

augmenter of metastasis in a murine melanoma model¹¹. It has also been implicated to be either pro- or anti-apoptotic. For example, on the one hand, mouse embryonic fibroblasts derived from *Atf3*^{-/-} mice are partially protected from stress-induced apoptosis¹². On the other hand, ATF3 protects endothelial cells from tumour-necrosis factor (TNF)-induced apoptosis by decreasing the transcription of p53 (ref. 13).

ATF3 regulatory complex and transcriptional targets

In order to determine the role of ATF3 in TLR signalling we first needed to identify the components of the putative ATF3 regulatory complex. Our strategy was to use computational tools to predict members of the complex, and then to validate the prediction using biological methods. Potential ATF3-interacting proteins were identified using a protein–protein interaction map displayed in Cytoscape (<http://www.cytoscape.org/>), a network analysis and visualization tool^{4,15} (Fig. 2a; the nodes (circles) represent proteins and the edges (lines) represent direct interactions between the proteins). The protein–protein interactions visualized in Cytoscape are largely curated from the literature, and, as such, they represent possible interactions. Notably, ATF3 is predicted to interact with two major transcriptional complexes known to be involved in TLR signalling: NF- κ B (p50) and AP1 (Jun, Junb, Jund1 and Fos) (Fig. 2a). We chose Rel as a surrogate for NF- κ B as it co-segregated with ATF3 in cluster 6. Jun and Fos are also members of cluster 6 and they were therefore chosen to represent AP1. Again we used MotifMogul, this time to identify enriched ATF3, NF- κ B and AP1 binding site densities and their proximity to each other and to the transcriptional start sites of potential target genes in cluster 2 (that is, the cluster of genes that was activated immediately after the induction of the transcription factors in cluster 6). Using these constraints we identified 30 target genes that contained putative ATF3 sites within 100 base pairs (bp) of NF- κ B binding sites and located within 500 bp of the transcriptional start site (Supplementary Data 3 and 4). Of this subset, IL-6 and IL-12b were chosen for further investigation because of their biological relevance (Fig. 2b).

ATF3 kinetic model in LPS-induced gene transcription

We then validated the predicted ATF3 and Rel binding sites on the regulatory regions of the genes encoding IL-6 and IL-12b using chromatin immunoprecipitation (ChIP) (Fig. 3a). Rel bound to the *Il6* and *Il12b* promoters within 1 h of LPS stimulation, and this binding declined after 2 h. ATF3 binding occurred more slowly, with maximal binding at 4 h. In contrast to Rel, ATF3 binding did not decline, and remained constant at 6 h (Fig. 3a). The binding of ATF3 and Rel to their cognate promoters mirrored the nuclear concentration of these transcription factors (Fig. 3b).

To characterize further the function of ATF3 in the TLR pathway we searched for functionally predictive kinetic relationships between the binding of Rel and ATF3 to the *Il6* promoter and *Il6* expression. Using a generalized multivariate regression model^{16,17}, *Il6* transcription levels are determined by the amount of promoter-bound ATF3 and Rel; these combine additively with weight coefficients β_{Rel} and β_{ATF3} , the numerical values of which can be estimated by fitting the model to the data (see Methods). These parameters describe the relative influence of each transcription factor in determining *Il6* expression, and are estimated from the *Il6* expression and transcription factor occupancy on the *Il6* promoter. A positive coefficient implies that the transcription factor has the role of an activator, whereas a negative coefficient implies repression. Fitting the parameters yields $\beta_{\text{Rel}} = 7.8$ and $\beta_{\text{ATF3}} = -4.9$, suggesting that Rel and ATF3 are both important in determining *Il6* expression levels, and that Rel is a transcriptional activator (consistent with current knowledge) whereas ATF3 is a negative regulator of *Il6* expression (Supplementary Data 5a). To explore further this prediction we simulated *Atf3*-null conditions by removing the corresponding term in the kinetic equation (Fig. 3c). The model predicts that *Il6* expression is substantially increased in *Atf3*-null conditions, and, in contrast to the wild type, the transcription of *Il6* is not inhibited, but continues to rise (Fig. 3c). An exhaustive search of the biologically plausible range of values for these parameters revealed only one possible answer: that ATF3 is a negative regulator of LPS-induced *Il6* expression (Fig. 3d). Similar conclusions were derived from analysis of the *Il12b* gene with

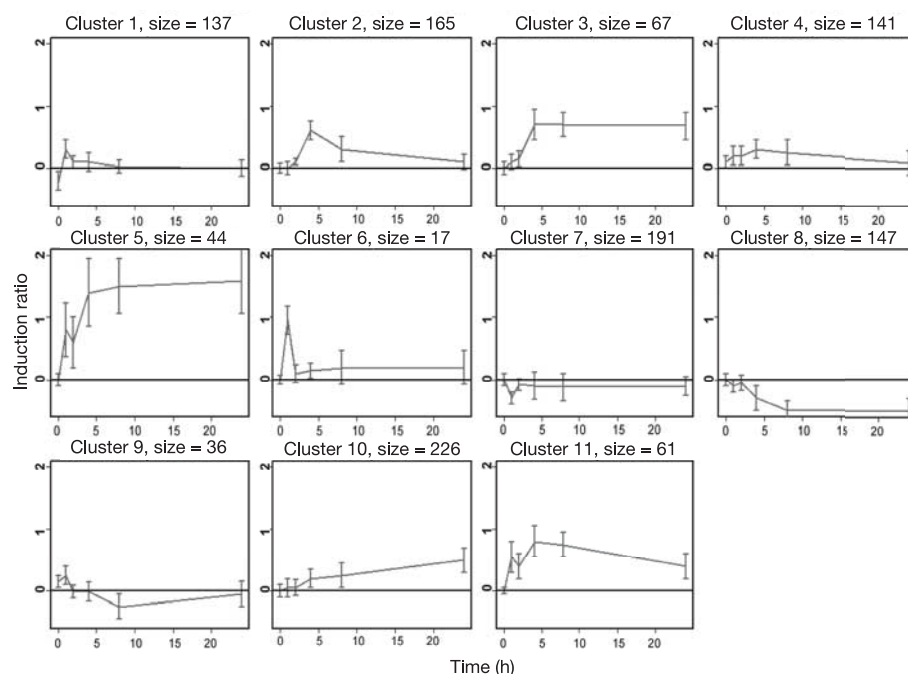


Figure 1 | Macrophage genes regulated by LPS form distinct kinetic clusters. Macrophages were stimulated with 10 ng ml⁻¹ LPS and RNA was isolated at the indicated times and subjected to microarray analysis. Genes were clustered by their kinetic profiles using a *k*-means algorithm using

average log₁₀ values of normalized gene expression ratios. Data represent the average of three independent experimental values $\pm$ standard error. The transcription factor constituents of cluster 6, which show an early peak of mRNA expression, are ATF3, Btg2, c-Jun, c-Fos, Egr1, Egr2, Ier2 and c-Rel.

fitting parameters of $\beta_{\text{Rel}} = 18.5$ and $\beta_{\text{Atf3}} = -9.6$ (Supplementary Data 5, Figs S51 and S52).

Biological effects of ATF3

To clarify the function of ATF3 in TLR4 signalling and to test our predictions, we obtained *Atf3*^{-/-} mice for further study¹⁴. These mice develop normally and show no overt immunological phenotype in specific pathogen-free conditions. Quantitative real-time polymerase chain reaction with reverse transcription (RT-PCR) revealed a substantial increase in LPS-induced *Il6* and *Il12b* mRNA levels in *Atf3*^{-/-} bone-marrow derived macrophages (BMMs) (Fig. 4a). LPS induction of *iNOS* and *Tnf* mRNA was similarly enhanced in *Atf3*^{-/-} mice. However, mRNA levels encoding MIP2 were unaffected in *Atf3*^{-/-} mice, demonstrating selectivity of the ATF3 effect (Fig. 4a). The increases in mRNA were mirrored by cytokine secretion and NO production (Fig. 4b). Similar effects were noted in resident peritoneal macrophages as well as in liver (Supplementary Data 2, Figs 2S2 and 2S3). The inhibitory role of ATF3 in LPS-induced responses in BMMs in culture was even more pronounced *in vivo*. Circulating levels of IL-6

and IL-12b were increased more than 10-fold in *Atf3*^{-/-} mice receiving LPS intraperitoneally (Fig. 4c) compared with wild-type counterparts. Serum TNF levels were similarly increased (Supplementary Data 2, Fig. 2S4). Survival in this *in vivo* endotoxemic shock model was also substantially altered: whereas 100% of *Atf3*^{-/-}

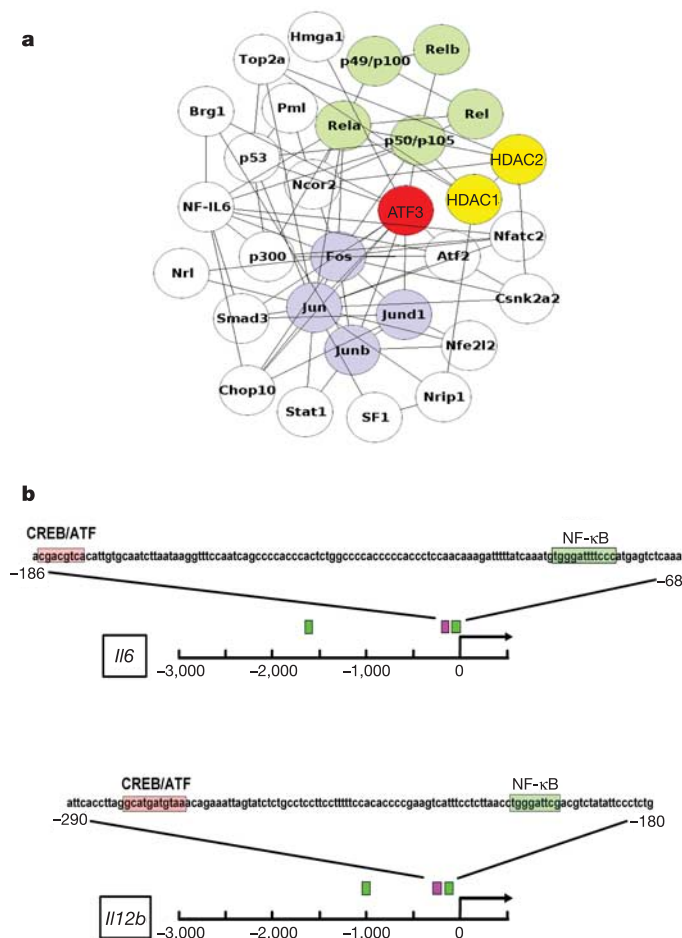


Figure 2 | Predicting ATF3 target genes using protein interaction network and promoter analysis. **a**, ATF3-associated transcription factors are determined from the known transcription factor protein–protein interaction network using Cytoscape. ATF3 (red) is predicted to interact with a number of transcription factors, including members of the AP1 (light blue) and NF-κB (light green) transcription factor complexes. **b**, Computational analysis of the regulatory elements of *Il6* and *Il12b*. ATF3/CREB and NF-κB binding sites were identified by scanning the 5' promoter regions of the *Il6* and *Il12b* genes with MotifMogul. Each coloured block represents a match: red blocks are ATF/CREB sites; green blocks are NF-κB sites. The numbers on the axis refer to bases 5' upstream from transcription start site.

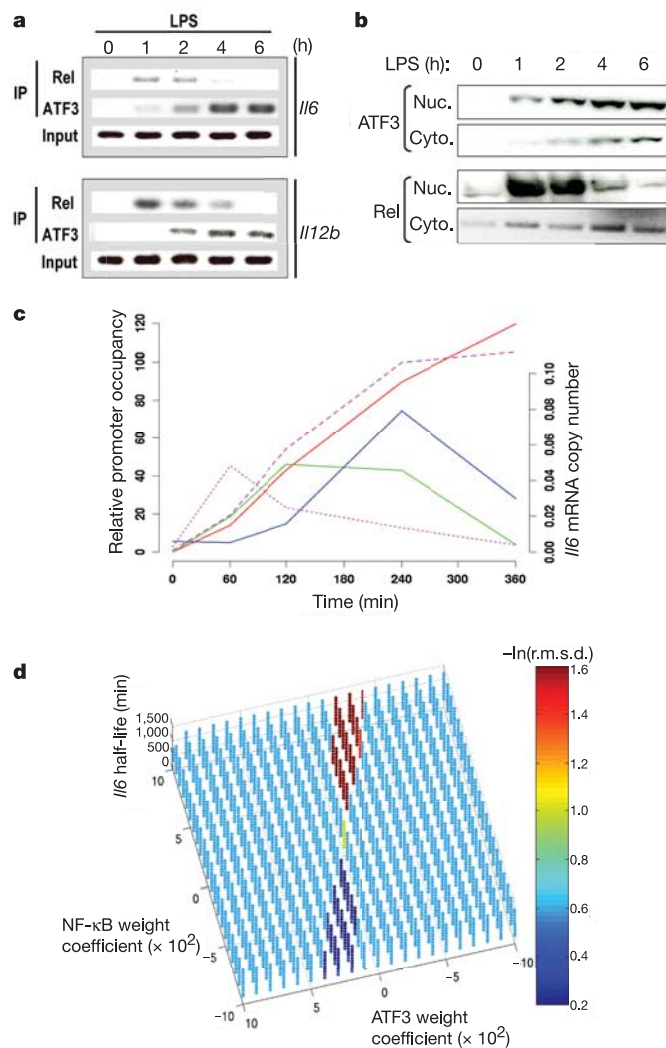


Figure 3 | Kinetic analysis predicts a negative regulatory role for ATF3 in LPS-stimulated responses. **a**, Temporal recruitment of ATF3 and Rel to the *Il6* and *Il12b* promoters. Macrophages from wild-type mice were stimulated with 10 ng ml⁻¹ LPS for the indicated times. ChIP assays were performed. DNA from input or immunoprecipitated (IP) fractions was measured by PCR amplification of specific promoter sequences. **b**, Protein expression and localization of ATF3 and Rel. Macrophages from wild-type mice were stimulated with 10 ng ml⁻¹ LPS for the indicated times. Nuclear and cytosolic extracts were analysed by immunoblotting. **c**, Kinetics of *Il6* regulation by Rel and ATF3. Kinetic transcription factor binding data for Rel and ATF3 in LPS-stimulated macrophages was determined from chromatin immunoprecipitation (Rel (dotted purple line) and ATF3 (dashed purple line)), and plotted together with *Il6* mRNA levels as determined by real-time PCR (blue solid line). Two parameters were optimized, $\beta_{\text{Rel}} = 7.8$ and $\beta_{\text{Atf3}} = -4.9$, representing the relative contributions of Rel and ATF3 levels in determining *Il6* transcription. Shown by the green solid line are the predicted *Il6* levels derived from the kinetic model. Omitting the ATF3 term in the model gives a prediction of *Il6* levels in an *Atf3* knockout (red solid line). **d**, Visualization of the parameter space of the kinetic model of *Il6* regulation. The space of biologically plausible parameter values was visualized as a three dimensional grid. The fit is best in the red region to the top and right of the plot, and worst in the dark blue region (to the bottom left of the figure).

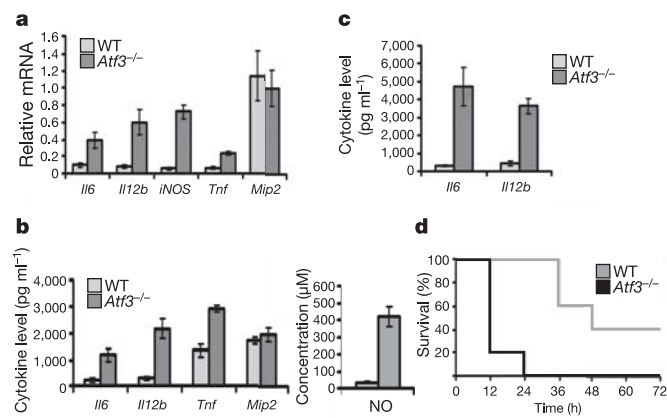


Figure 4 | Role of ATF3 in *in vitro* and *in vivo* LPS responses. **a**, BMMs were stimulated with 10 ng ml^{-1} LPS for 4 h and the indicated cytokines were measured by quantitative real-time RT-PCR analysis. Data represent the average of three independent experimental values $\pm$ standard error. WT, wild-type mice (C57BL/6). **b**, BMMs were stimulated with 10 ng ml^{-1} LPS for 4 h and cytokine levels determined by ELISA (left) and nitric oxide determined by Griess assay (right). Data represent the average of three independent experimental values $\pm$ standard error. **c**, Wild-type or *Atf3*^{-/-} mice were injected intraperitoneally with LPS (20 mg kg^{-1}), serum was drawn at 12 h and analysed by ELISA. Data represent the average of three values $\pm$ standard error. **d**, Age-matched wild-type and *Atf3*^{-/-} mice ($n = 10$) were injected intraperitoneally with LPS (20 mg kg^{-1}) and then monitored for survival.

mice had succumbed by 24 h after LPS administration, none of the wild-type mice had expired at this early time point (Fig. 4d).

Functional mechanism of ATF3

We next explored the mechanism by which ATF3 inhibits LPS-induced transcription. Chromatin remodelling regulates transcription by allowing, or preventing, access of transcription factors to their cognate binding sites. The protein-protein interaction map demonstrates connectivity of ATF3 with histone deacetylase (HDAC; yellow nodes) via the NF- κ B complex (Fig. 2a). HDAC has a critical role in chromatin remodelling by participating in the acetylation/deacetylation cycle of histones: acetylation relaxes chromatin structure thereby allowing access of transcription factors to DNA, conversely, deacetylation alters chromatin structure to limit transcription factor access. LPS treatment of macrophages resulted in a substantial increase in HDAC activity that could be co-precipitated with ATF3 (Fig. 5a, left panel), and western blotting demonstrated the presence of HDAC1 in the immunoprecipitated complex (Fig. 5a, right panel). We assessed the effect of ATF3-dependent chromatin remodelling on the *Il6* and *Il12b* promoters by determining histone acetylation status using ChIP. Acetylated histone H4 (H4ac), an indicator of open chromatin structure, bound to the *Il6* promoter within 1 h of LPS treatment (Fig. 5b). The kinetics of the association of H4ac with the *Il6* promoter mirrored the binding of Rel to this promoter (compare Figs 5b and 3a), strongly suggesting that the open chromatin structure permits access of Rel to its DNA binding site. After 2 h the amount of promoter-bound H4ac decreased, with a temporal response similar to the decrease in Rel binding (Figs 5b and 3a). These kinetics mirror the association of ATF3 with the *Il6* promoter, and the observation that deacetylation of histone H4 does not occur in *Atf3*^{-/-} macrophages strongly suggests that ATF3-associated HDAC1 is responsible for this deacetylation (Figs 5a, b and 3a). Similar results were obtained for the association of H4ac with the *Il12b* promoter, and for acetylated histone H3 on both promoters (data not shown). The observation that the HDAC inhibitor trichostatin A increased LPS-induced transcription of *Il6* and *Il12b* in a similar manner to that seen in *Atf3*^{-/-} macrophages further suggested that ATF3 inhibits transcription via chromatin remodelling (Fig. 5c).

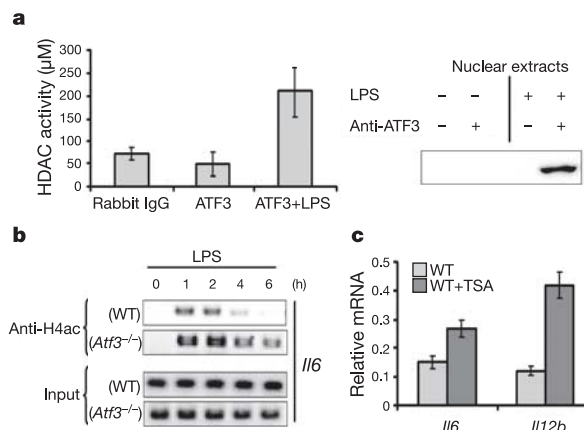


Figure 5 | Role of HDAC in ATF3-mediated gene regulation. **a**, Left panel: BMMs from wild-type mice were stimulated with 10 ng ml^{-1} LPS for 6 h and immunoprecipitation with relevant antibody and associated HDAC activity measured by colorimetric assay. Data represent the average of three independent experimental values $\pm$ standard error. Right panel: complex formation of ATF3 and HDAC1 *in vivo*. BMMs from wild-type mice were stimulated with 10 ng ml^{-1} LPS for 6 h and nuclear extracts were isolated and immunoprecipitated with anti-ATF3 antibody and immunoblotted with anti-HDAC1. **b**, Dynamics of H4 acetylation on the *Il6* promoter. Macrophages from wild-type or *Atf3*^{-/-} mice were stimulated with 10 ng ml^{-1} LPS for the indicated times. ChIP assays were performed using anti-acetyl H4 antibody as described in the Methods. **c**, Treatment with HDAC inhibitor potentiates macrophage cytokine gene transcription. BMMs from wild-type mice (C57BL/6) were treated with the HDAC inhibitor trichostatin A (TSA, 100 ng ml^{-1}) and 10 ng ml^{-1} LPS for 4 h and the cytokines measured by quantitative real-time RT-PCR analysis. Data represent the average of three independent experimental values $\pm$ standard error.

Taken together, the data suggest that LPS-induced acetylation of histones opens chromatin structure to allow access of Rel to the *Il6* and *Il12b* promoters and activation of transcription. ATF3 subsequently binds to these promoters, and ATF3-associated HDAC1 deacetylates histones, resulting in the closure of chromatin and the inhibition of transcription. This mechanism is indirectly supported by a recent publication demonstrating that ATF3 Δ Zip2, a splice variant of ATF3 lacking DNA-binding activity, sensitizes cells to apoptosis by sequestering CREB-binding protein-p300 (ref. 18). Because CREB-binding protein-p300 form a scaffold for histone acetyltransferase (HAT)¹⁹, their sequestration would indirectly result in a lack of histone acetylation. Thus, it is possible that ATF3 may modify transcription by two independent mechanisms involving histone acetylation: first, in a DNA-dependent manner by directly recruiting a histone deacetylase to promoters containing an ATF3 binding site, and second, in a DNA-independent manner by ATF3 Δ Zip2, which sequesters the HAT complex, thus indirectly preventing histone acetylation.

ATF3 transcriptional network

Our data demonstrate that ATF3 and Rel jointly regulate LPS-induced *Il6*, *Il12b* and *iNOS* transcription. We used ChIP-to-chip analysis (see below) to examine whether ATF3 binds to the *cis*-regulatory elements of the remaining 27 genes (listed in Supplementary Data 3, Table S3.1) that were computationally predicted to be its targets. DNA segments bound by ATF3 were chromatin immunoprecipitated (ChIP) and hybridized to a DNA microarray (chip) containing 25-mer oligonucleotide probes tiling the proximal promoters of selected genes at a density of approximately every 20–30 bases. Eleven genes with significantly increased ATF3 binding were identified (Supplementary Data 8, Fig. S8a–c). We cross-validated this set against a set of 15 genes that are upregulated in *Atf3*^{-/-} macrophages, obtained using Affymetrix microarrays. As shown in Fig. 6a, 11 genes are significantly

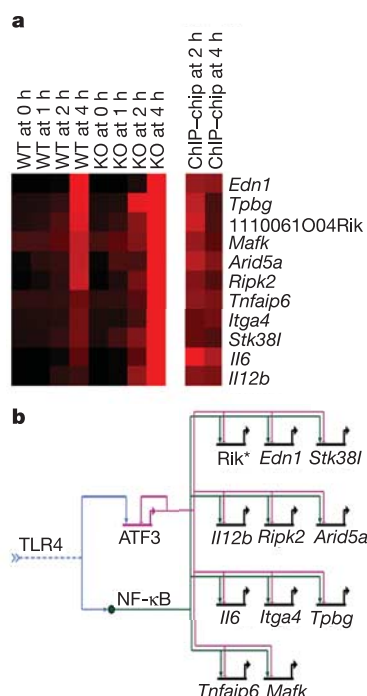


Figure 6 | The transcriptional regulatory network of ATF3. **a**, Correlation of the ATF3 DNA binding profiles with gene expression patterns. Macrophages from wild-type or *Atf3*^{-/-} mice were stimulated with 10 ng ml⁻¹ LPS and RNA was isolated at the indicated times and subjected to microarray analysis. ChIP binding data (at 2 h and 4 h) were derived from wild-type macrophages stimulated with 10 ng ml⁻¹ LPS. Heat map colour for expression arrays is based on normalized fold changes among eight conditions on each gene (from no expression (black) through to high expression (bright red)). ChIP-to-chip results are based on normalized peak ratios of identified segments for the 11 genes. **b**, Proposed ATF3 regulatory network in LPS-stimulated macrophages visualized using Biatapestry software (<http://www.biatapestry.org>). ATF3 and NF-κB are both activated by TLR4 signalling and jointly regulate the transcription of a battery of downstream genes. *Rik**, 1110061O04Rik.

upregulated in *Atf3*^{-/-} macrophages and directly bound by ATF3. Taken together, these data suggest that ATF3 is a transcriptional repressor of the genes in Fig. 6a. Figure 6b is a network diagram incorporating the findings reported in this paper. Thus, ATF3 and NF-κB are both induced by TLR4 signalling and jointly regulate a battery of downstream genes. Because ATF3 is itself induced by LPS, it seems to regulate TLR-stimulated inflammatory responses as part of a negative feedback loop.

METHODS

Mouse bone-marrow-derived macrophages. *Atf3*^{-/-} mice in the C57BL/6 background were generated by backcrossing *Atf3*^{-/-} mice in the 129SVJ background to C57BL/6 mice, ten times¹⁴. Bone marrow was collected from femurs in complete RPMI containing 10% heat-inactivated FCS (Hyclone Laboratories), 100 U ml⁻¹ penicillin, 100 μg ml⁻¹ streptomycin, 2 mM L-glutamine and 50 ng ml⁻¹ rmM-CSF (R&D Systems). BMMs were stimulated with high purity LPS (10 ng ml⁻¹) for the indicated times. As a model of LPS-induced septic shock, mice were injected intraperitoneally with 20 mg kg⁻¹ LPS. The University of Washington and Institute for Systems Biology's Institutional Animal Care and Use Committees approved all animal protocols.

Affymetrix GeneChip analysis and quantitative real-time PCR. Total RNA was isolated using Trizol (Invitrogen) and analysed by real-time PCR with probes (IDT) labelled with 5' FAM and 3' TAMRA. Data were normalized by the level of EF1a expression in individual samples. Primer sequences are available upon request. The Affymetrix protocol was essentially as described. cRNA was hybridized for 16 h to Affymetrix MG-U74Av2 arrays (Affymetrix), which contain ~45,101 mouse probe sets, half of which correspond to expressed sequence tags (ESTs) and half to characterized genes. Normalization was performed using GeneChip robust multi-array analysis, followed by GC-robust

multi-array average (GC-RMA) normalization. Identification of significantly perturbed genes was done using significance analysis of microarrays. The false positive rate was 0.1%.

Immunoblotting. Whole-cell and nuclear/cytosolic extracts of BMMs were prepared as previously described²⁰. Antibodies used for immunoblotting were purchased as indicated: ATF3 (Santa Cruz), Rel (Santa Cruz), acetylated histone H3 (Upstate) and acetylated histone H4 (Upstate).

Chromatin immunoprecipitation assay. After stimulation, BMMs were fixed in formaldehyde and ChIP assay was performed. The antibodies used for immunoprecipitation were as listed above. The purified DNA was analysed by PCR using primers specific for the *Il6* and *Il12b* promoters. The PCR products were visualized on an ethidium bromide gel. The levels of promoter-bound ATF3 and Rel were determined by band densitometry for use in kinetic modelling. A detailed protocol is noted in Supplementary Data 6.

ELISA. Cytokine release were determined with a sandwich enzyme-linked immunosorbent assay (ELISA) technique (Duoset; R&D Systems) using the manufacturer's recommended protocol.

MotifMogul. We scanned the promoter regions of genes from -3,000 bp to +500 bp of the transcriptional starting site (NCBI m33 mouse assembly). Three scanning methods—MotifLocator²¹, MotifScanner²¹ and Clover²²—were applied using known TRANSAC motifs to search ATF/CREB, NF-κB and AP1 binding sites (Professional version 8.3)²³. Because ATF binding sites have strong overlap with CREB binding sites, we scanned using a combined matrix set including three ATF matrices and nine CREB matrices. MotifMogul is used to integrate the three scanning methods, and default parameters of each individual algorithm were applied. Only statistically significant sites, as determined against a randomly derived DNA sequence, are filtered and examined. Matrix matches were visualized using the Apollo genome annotation viewer²⁴.

Cytoscape. Cytoscape, our network visualization program, overlays cDNA expression profiles on a matrix of protein–protein and protein–DNA interaction networks. Our Cytoscape database contains 5,200 known protein–protein and protein–DNA interactions, and 17,600 relationships between these proteins (mostly curated from the literature). For a comprehensive description of these programs, see <http://www.systemsbiology.org>.

Kinetic model. We used a generalized multivariate regression model^{16,17} to model the kinetics of transcriptional regulation of *Il6* or *Il12b* by ATF3 and Rel:

$$\tau \frac{d(Il6)}{dt} = -Il6 + g(\beta_{Rel}Rel + \beta_{ATF3}ATF3) \quad (1)$$

In equation (1), the term *Il6* represents the level of *Il6* mRNA expression as determined by real-time PCR, whereas the terms *Rel* and *ATF3* represent the levels of Rel and ATF3 bound to the *Il6* promoter, as determined by ChIP. The weight parameters β_{Rel} and β_{ATF3} , estimated from the data, are the relative contributions of Rel and ATF3 in determining *Il6* transcription according to the model. The function *g* is a sigmoidal transfer function¹⁷, incorporating lower and upper bounds on *Il6* expression. The parameter τ was fixed for consistency with $T^{1/2} = 600$ min, a representative mRNA half-life in mammalian cells²⁵. The kinetic equation was discretized at the measured time intervals¹⁷, and the optimal parameters were found to be $\beta_{Rel} = 7.8$ and $\beta_{ATF3} = -4.9$, by least squares regression. We verified that the signs of the two terms were preserved for a range of mRNA half-lives. Results for *Il12b* were obtained by a similar process.

Assay for HDAC activity. HDAC activity was measured by a non-isotopic assay (Biomol) using a chromagen linked acetylated substrate according to the manufacturer's protocol.

ChIP-to-chip analysis. We designed a custom oligonucleotide array containing densely tiled 25-mer oligonucleotide sequences designed to interrogate genes that were identified to be of probable importance for immune activation. For ChIP-to-chip binding analysis, we used an approach similar to previous publications^{26,27}, including quantile normalization²⁶ and median-based filtering with identification of ATF3 binding sites²⁷.

Received 28 February; accepted 29 March 2006.

- Kawai, T. & Akira, S. Pathogen recognition with Toll-like receptors. *Curr. Opin. Immunol.* **17**, 338–344 (2005).
- Aderem, A. Systems biology: its practice and challenges. *Cell* **121**, 511–513 (2005).
- Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
- Suthram, S., Sittler, T. & Ideker, T. The plasmodium protein network diverges from those of other eukaryotes. *Nature* **438**, 108–112 (2005).
- Royet, J., Reichhart, J. M. & Hoffmann, J. A. Sensing and signaling during infection in *Drosophila*. *Curr. Opin. Immunol.* **17**, 11–17 (2005).
- Takeda, K. & Akira, S. TLR signaling pathways. *Semin. Immunol.* **16**, 3–9 (2004).
- Nau, G. J. et al. Human macrophage activation programs induced by bacterial pathogens. *Proc. Natl Acad. Sci. USA* **99**, 1503–1508 (2002).
- Werner, T. Target gene identification from expression array data by promoter analysis. *Biomol. Eng.* **17**, 87–94 (2001).

9. Mayr, B. & Montminy, M. Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nature Rev. Mol. Cell Biol.* **2**, 599–609 (2001).
10. Yan, C., Lu, D., Hai, T. & Boyd, D. D. Activating transcription factor 3, a stress sensor, activates p53 by blocking its ubiquitination. *EMBO J.* **24**, 2425–2435 (2005).
11. Ishiguro, T. *et al.* Identification of genes differentially expressed in B16 murine melanoma sublines with different metastatic potentials. *Cancer Res.* **56**, 875–879 (1996).
12. Ishiguro, T., Nagawa, H., Naito, M. & Tsuruo, T. Inhibitory effect of ATF3 antisense oligonucleotide on ectopic growth of HT29 human colon cancer cells. *Jpn. J. Cancer Res.* **91**, 833–836 (2000).
13. Kawauchi, J. *et al.* Transcriptional repressor activating transcription factor 3 protects human umbilical vein endothelial cells from tumor necrosis factor- α -induced apoptosis through down-regulation of p53 transcription. *J. Biol. Chem.* **277**, 39025–39034 (2002).
14. Hartman, M. G. *et al.* Role for activating transcription factor 3 in stress-induced beta-cell apoptosis. *Mol. Cell. Biol.* **24**, 5721–5732 (2004).
15. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
16. Thorsson, V. T., Hörnquist, M., Siegel, A. F. & Hood, L. Reverse engineering galactose regulation in yeast through model selection. *Stat. Appl. Gen. Mol. Biol.* **4**, 1–24 (2005).
17. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data-sets *de novo*. *Genome Biol.* **7**, R36–R42 (2006).
18. Hua, B. *et al.* A splice variant of stress response gene ATF3 counteracts NF- κ B-dependent anti-apoptosis through inhibiting recruitment of CREB-binding protein/p300 coactivator. *J. Biol. Chem.* **281**, 1620–1629 (2006).
19. Furumatsu, T. *et al.* Sox9 and p300 cooperatively regulate chromatin-mediated transcription. *J. Biol. Chem.* **280**, 35203–35208 (2005).
20. Gilchrist, M., McCauley, S. D. & Befus, A. D. Expression, localization, and regulation of NOS in human mast cell lines: effects on leukotriene production. *Blood* **104**, 462–469 (2004).
21. Thijs, G. *et al.* INCLUSIVE: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics* **18**, 331–332 (2002).
22. Frith, M. C. *et al.* Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* **32**, 1372–1381 (2004).
23. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
24. Lewis, S. E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3**, research0082 (2002).
25. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).
26. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
27. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge A. Ozinsky, I. Shmulevich, W. Longabaugh and L. Hood for discussions. We thank A. Nachman, A. Clark and C. Baldwin for technical assistance. This work was supported by a Fellowship from the Alberta Heritage Foundation for Medical Research (to M.G.) and the NIH (to A.A.)

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.A. (aaderem@systemsbiology.org).

A mutation in Orai1 causes immune deficiency by abrogating CRAC channel function

Stefan Feske^{1,2,*}, Yousang Gwack^{1,3,*}, Murali Prakriya⁴, Sonal Srikanth^{1,3}, Sven-Holger Puppel¹, Bogdan Tanasa¹, Patrick G. Hogan¹, Richard S. Lewis⁵, Mark Daly^{6,7} & Anjana Rao^{1,3}

Antigen stimulation of immune cells triggers Ca^{2+} entry through Ca^{2+} release-activated Ca^{2+} (CRAC) channels, promoting the immune response to pathogens by activating the transcription factor NFAT. We have previously shown that cells from patients with one form of hereditary severe combined immune deficiency (SCID) syndrome are defective in store-operated Ca^{2+} entry and CRAC channel function. Here we identify the genetic defect in these patients, using a combination of two unbiased genome-wide approaches: a modified linkage analysis with single-nucleotide polymorphism arrays, and a *Drosophila* RNA interference screen designed to identify regulators of store-operated Ca^{2+} entry and NFAT nuclear import. Both approaches converged on a novel protein that we call Orai1, which contains four putative transmembrane segments. The SCID patients are homozygous for a single missense mutation in *ORAI1*, and expression of wild-type Orai1 in SCID T cells restores store-operated Ca^{2+} influx and the CRAC current (I_{CRAC}). We propose that Orai1 is an essential component or regulator of the CRAC channel complex.

Ca^{2+} is an essential second messenger in almost all cell types. Sustained Ca^{2+} influx across the plasma membrane is crucial for lymphocyte activation and the adaptive immune response^{1–4}. Antigen recognition by T and B lymphocytes triggers phospholipase $\text{C}\gamma$ activation, inositol-1,4,5-triphosphate (IP_3) generation and the release of Ca^{2+} from endoplasmic reticulum (ER) stores. Depletion of ER Ca^{2+} stores opens CRAC channels, a class of ‘store-operated’ Ca^{2+} channels with high selectivity for Ca^{2+} over monovalent cations, low single-channel conductance (<1 pS) and an inwardly rectifying current–voltage (I – V) relationship^{1,5–8}. One of the main Ca^{2+} -regulated transcription factors is NFAT, a family of heavily phosphorylated proteins that reside in the cytoplasm of resting cells^{2,3}. Sustained Ca^{2+} influx results in NFAT dephosphorylation by the calmodulin-dependent protein phosphatase calcineurin and promotes NFAT translocation to the nucleus.

CRAC channels are a principal pathway for Ca^{2+} influx in T cells^{1–4,9,10}. T cells from two patients with hereditary severe combined immune deficiency (SCID) syndrome, who presented as infants with a marked propensity for fungal and viral infections^{11,12}, had a primary defect in store-operated Ca^{2+} entry and CRAC channel function^{9,13}. The SCID T cells also showed a severe impairment in NFAT-dependent gene activation^{13–15}, highlighting the importance of CRAC channel function for lymphocyte activation and immune defence.

Although the pharmacological and electrophysiological properties of the CRAC channel have been described in detail^{1,5–7,16}, its molecular identity is unknown. Several members of the transient receptor potential (TRP) family of ion channels have been proposed as candidates^{17–20}, but none of them has all the biophysical properties

expected of the CRAC channel^{7,8,21,22}. Sequence analyses and complementation studies in the SCID patients’ cells⁹ failed to establish a causative role for several TRP family members (TRPC3, TRPV5 and TRPV6) or for the EF-hand-containing membrane proteins STIM1 and STIM2, which couple store depletion to CRAC channel opening by sensing the filling state of ER Ca^{2+} stores^{23–26}.

Here we identify a novel protein, that we designate Orai1, which is crucial for store-operated Ca^{2+} entry and CRAC channel function. We identified Orai1 using two unbiased genetic approaches: a modified linkage analysis to identify the gene mutated in the SCID patients, and a genome-wide RNA interference (RNAi) screen in *Drosophila* to identify regulators of store-operated Ca^{2+} entry and NFAT nuclear import. The combination of these two approaches pinpointed a single candidate gene. RNAi-mediated depletion of *Drosophila* Orai (dOrai) markedly diminished store-operated Ca^{2+} entry; likewise, the Ca^{2+} influx defect in the SCID patients was traced to a point mutation in human *ORAI1*. Complementation of SCID T cells and fibroblasts with wild-type Orai1 reconstituted store-operated Ca^{2+} influx and the CRAC channel current (I_{CRAC}), establishing a critical role for Orai1 in T-cell function and the *in vivo* immune response. We propose that Orai1 is a subunit or key regulator of the CRAC channel complex.

Phenotypic identification of heterozygous disease carriers

The two SCID patients in our study were born to consanguineous parents, suggesting a recessive mode of inheritance as neither the patients’ parents nor any other members of their extended family showed clinical symptoms of immunodeficiency (Fig. 1a). In the

¹The CBR Institute for Biomedical Research, and the Departments of ²Pediatrics and ³Pathology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA. ⁴Department of Molecular Pharmacology and Biological Chemistry, Northwestern University, Feinberg School of Medicine, Chicago, Illinois 60611, USA. ⁵Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. ⁶Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁷Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

*These authors contributed equally to this work.

presence of 2 mM extracellular Ca^{2+} ($[\text{Ca}^{2+}]_o$), T cells derived from the parents of the SCID patients showed almost normal store-operated Ca^{2+} entry¹³. To unmask a potential defect in Ca^{2+} entry in the parental T cells, we limited the driving force for Ca^{2+} entry by reducing $[\text{Ca}^{2+}]_o$ to 0.2–0.5 mM, and estimated the rate of Ca^{2+} influx indirectly by measuring the initial rate of change in intracellular Ca^{2+} concentration ($d[\text{Ca}^{2+}]_i/dt$) in cells treated with thapsigargin, an inhibitor of the SERCA Ca^{2+} pump. Under these conditions, peak $[\text{Ca}^{2+}]_i$ levels and initial rates of $[\text{Ca}^{2+}]_i$ increase in T cells from the parents were $\sim 50\%$ of those observed in wild-

type control T cells (Fig. 1b). This finding is consistent with a gene-dosage effect, in which the parents are heterozygous carriers of the causal mutation for SCID.

We used this assay to identify, in a more extended pedigree, other potentially heterozygous carriers of the mutation responsible for the SCID phenotype. Blood samples were obtained from 19 additional family members (Fig. 1a), T-cell lines were generated, and Ca^{2+} entry measurements were made. T cells from 13 family members consistently showed reduced peak $[\text{Ca}^{2+}]_i$ levels and decreased initial rates of $[\text{Ca}^{2+}]_i$ increase, compared to cells from eight other family members and unrelated controls (Fig. 1c and data not shown). On the basis of these data, a cutoff value of 2 nM s^{-1} for the rate of $[\text{Ca}^{2+}]_i$ increase was used to distinguish potential heterozygous disease carriers from unaffected individuals (Fig. 1c). With this cutoff, the distribution of putative heterozygous carriers within the family is compatible with an autosomal recessive mode of inheritance of the clinical disease (Fig. 1a).

Linkage mapping by genome-wide SNP array screen

Genomic DNA from the 23 members of the SCID family was used for genotyping using SNP mapping arrays. Data from the two patients, their parents, their unaffected brother and their grandparents were first evaluated assuming an autosomal recessive mode of inheritance based on the clinical phenotype (Pedigree A, grey shaded area in Fig. 1a). This analysis identified six regions with LOD ($\log_{10}$ of the odds ratio) scores of 1.5–1.9 (Fig. 2a, top panel). Although one of these regions is almost certain to harbour the defective gene, the LOD scores are significantly below the 3.0 value necessary to establish linkage, and so might be achieved by chance. We therefore implemented a second, completely independent analysis in which we assumed an autosomal dominant mode of inheritance, based on our ability to identify heterozygous carriers of the disease mutation by phenotypic analysis *in vitro* (Pedigree B, green box in Fig. 1a). The haplotype of twelve putatively heterozygous individuals was compared to that of the remaining eight homozygous healthy family members, identifying a unique region on chromosome 12q24 with a LOD score of ~ 3.8 , which overlapped with one of the six regions identified by the first approach (Fig. 2a, middle panel). This defines a $\sim 9.8\text{-Mb}$ candidate region with a highly significant cumulative LOD score of 5.7, representing odds of $\sim 500,000:1$ in favour of linkage (Fig. 2a, bottom panel and Fig. 2b). Genomic sequencing of six known genes in this region with a potential role in Ca^{2+} signalling or Ca^{2+} binding (*PLA2G1B*, *CABP1*, *P2RX4*, *P2RX7*, *CAMKK2* and *PITPNM2*) did not reveal any mutations in exons or immediately adjacent genomic regions, but did allow us to narrow down the candidate homozygous region to $\sim 6.5 \text{ Mb}$ containing ~ 74 genes, on the basis of several SNPs in *PITPNM2* for which the patients were heterozygous (Fig. 2c).

Drosophila olf186-F is a regulator of Ca^{2+} entry

In parallel with the positional cloning effort, we conducted a genome-wide RNAi screen^{27,28} for NFAT regulators in *Drosophila*, as an independent method of identifying components of the CRAC channel and the signalling pathway leading to CRAC channel activation. This screen takes advantage of the fact that although Ca^{2+} -regulated NFAT proteins are not represented in *Drosophila*, there is strong evolutionary conservation of the pathways that regulate NFAT shuttling between the cytoplasm and nucleus²⁹. For example, *Drosophila* S2 cells contain a store-operated Ca^{2+} channel that is very similar to the CRAC channel³⁰. *Drosophila* S2R+ cells stably expressing an NFAT-green fluorescent protein (GFP) fusion protein were incubated for four days with arrayed double-stranded (ds)RNAs against each of $\sim 21,000$ *Drosophila* genes to achieve knockdown of gene expression²⁹. The cells were stimulated for 10 min with thapsigargin to deplete Ca^{2+} stores, thus activating store-operated Ca^{2+} entry and nuclear translocation of NFAT-GFP, then fixed and photographed robotically.

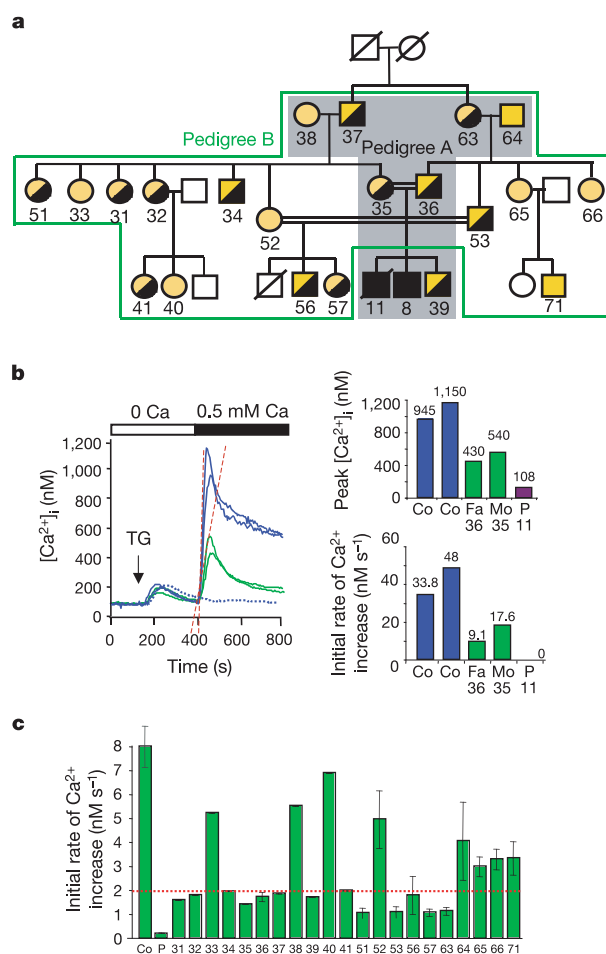


Figure 1 | Gene-dosage effect in store-operated Ca^{2+} entry. **a**, Pedigree of SCID patients with a defect in store-operated Ca^{2+} entry. Black boxes indicate patients with SCID (individuals 8 and 11), diagonal bars indicate deceased individuals, symbols consisting of two different colours indicate heterozygous disease carriers as determined by phenotypic analysis, and double horizontal bars indicate consanguineous marriages. DNA and lymphocytes were obtained for functional and genetic analysis from all individuals to whom a number is assigned. **b**, Reduced Ca^{2+} influx in T cells of both parents of the SCID patients (green) compared to wild-type controls (blue). Peak $[\text{Ca}^{2+}]_i$ (top right panel) and initial rates of $[\text{Ca}^{2+}]_i$ increase (bottom right panel) were measured in thapsigargin (TG)-stimulated T cells after re-addition of 0.5 mM extracellular Ca^{2+} . Dashed red lines indicate initial slopes of $[\text{Ca}^{2+}]_i$ increase. The dashed blue trace represents the Ca^{2+} response in T cells from SCID patients. **c**, Reduced store-operated Ca^{2+} entry in T cells from 13 out of 21 family members of the SCID patients identifies them as heterozygous disease carriers. Shown are averages of the initial rates of $[\text{Ca}^{2+}]_i$ influx in the presence of 0.2 mM $[\text{Ca}^{2+}]_o$. Identification numbers correspond to individuals shown in **a**. Stars indicate heterozygous carriers with initial rates of Ca^{2+} influx less than or equal to 2 nM s^{-1} (dotted red line). Co, healthy control; P, SCID patient; Fa, father; Mo, mother. Error bars represent s.e.m. from three independent experiments (50–100 cells analysed per experiment).

Among the positive candidates for which depletion interfered with nuclear translocation and dephosphorylation of NFAT–GFP were *Drosophila Stim* (*dStim*) and *olf186-F* (hereafter designated *dOrai*, for reasons discussed below) (Fig. 3a, b). Knockdown of either *dStim* or *dOrai* completely inhibited thapsigargin-induced Ca^{2+} influx in S2R+ cells (Fig. 3c), but had no effect on the filling state of Ca^{2+} stores (Supplementary Fig. 1). These data confirm previous reports that dSTIM is essential for activation of store-operated Ca^{2+} entry and of a CRAC channel-like current in *Drosophila* cells²⁴, and identify *dOrai* as a second regulator of this process.

olf186-F has three human homologues, *FLJ14466* on chromosome 12, *C7orf19* on chromosome 7 and *MGC13024* on chromosome 16.

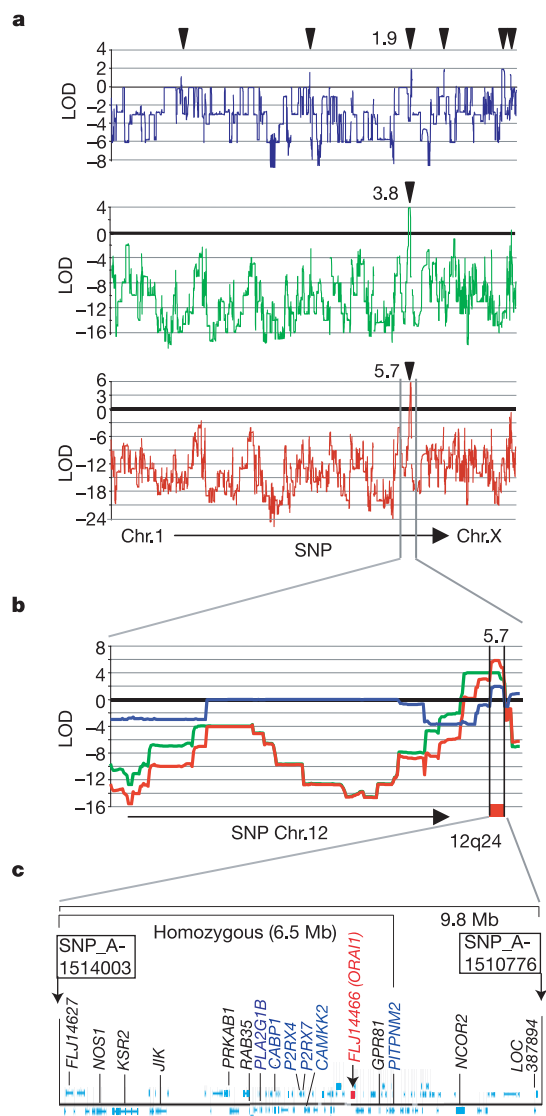


Figure 2 | A region on chromosome 12q24 is linked to the SCID gene defect. Genome-wide linkage analysis in 23 individuals from the SCID family. **a**, Multipoint parametric LOD scores were calculated for >10,000 SNPs. LOD scores derived from autosomal recessive (top) and dominant (middle) analyses were combined to yield a 'cumulative' LOD score (bottom) and plotted against the position of SNP markers in the genome (horizontal axis). Arrows indicate regions statistically linked to the SCID disease. **b**, A unique candidate gene region on 12q24 shows a highly significant cumulative LOD score (red trace) of 5.7. **c**, The ~6.5-Mb genomic interval on 12q24 contains *FLJ14466* (*ORAI1*). Genes sequenced in the SCID patients are shown in blue, other genes in black. The graphic representation of the candidate region was modified from NCBI MapViewer (<http://www.ncbi.nlm.nih.gov/mapview/>).

Notably, *FLJ14466* is located within the 6.5-Mb homozygous genomic region linked to the SCID mutation at 12q24 (Fig. 2c), and contains the causal mutation for the SCID syndrome as shown below. We have named the proteins encoded by *FLJ14466*, *C7orf19* and *MGC13024* as Orai1, Orai2 and Orai3, respectively. (In Greek mythology, the Orai are the keepers of the gates of heaven: Eunomia (Order or Harmony), Dike (Justice) and Eirene (Peace)^{31,32}.)

ORAI1 is mutated in the SCID patients

By sequencing genomic DNA from the 23 individuals (patients and their relatives) indicated in Fig. 1a, we found that the SCID defect is associated with a missense mutation in exon 1 of human *ORAI1* (Fig. 4a). The mutation is a C→T transition at position 271 of the coding sequence of *ORAI1* (position 444 of NM_032790), leading to replacement of a highly conserved arginine residue by tryptophan at position 91 of the protein (R91W) (Supplementary Fig. 2). All 13 phenotypically predicted heterozygous disease carriers (Fig. 1) were genotypically heterozygous for the mutation (C/T), whereas healthy controls and unaffected family members were homozygous for the wild-type allele (C/C) (Fig. 4a and data not shown).

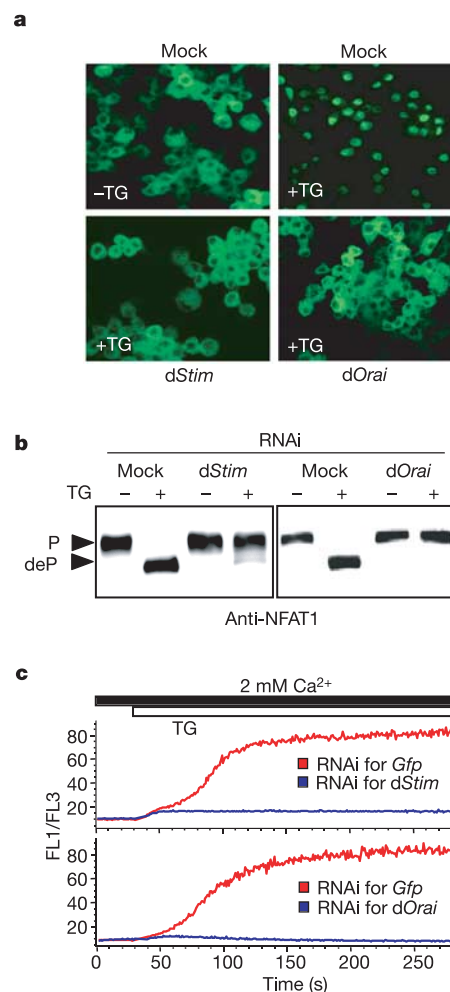


Figure 3 | *Drosophila* Orai regulates NFAT translocation and Ca^{2+} influx. S2R+ cells stably transfected with GFP–NFAT1(1–460) were incubated for four days with dsRNA against *dStim*, *dOrai* or an irrelevant DNA sequence (mock). **a**, **b**, Knockdown of *dOrai* or *dStim* prevents thapsigargin-induced nuclear translocation (**a**) and dephosphorylation (deP) (**b**) of NFAT–GFP. **c**, RNAi-mediated depletion of *dStim* or *dOrai* inhibits Ca^{2+} influx in S2R+ cells. Cells were loaded with Fluo-4 and Fura-Red and analysed for Ca^{2+} influx by flow cytometry after stimulation with thapsigargin (1 μM). Traces show changes in the emission ratio of the Ca^{2+} indicator dyes Fluo-4 (FL1) and Fura Red (FL3).

The C→T mutation is not an annotated SNP (in dbSNP Build 124), nor is it represented in the DNA from 270 individuals of mixed ethnic backgrounds assembled for the international HapMap project³³ (data not shown), suggesting that it is not a common sequence variant in the general population. The mutated residue is located at the beginning of the first of four potential transmembrane segments in Orai1, predicted by calculating the hydrophobicity of Orai1 using the Kyte–Doolittle method³⁴ (Fig. 4b). We showed that Orai1 is localized at or near the plasma membrane by expressing amino- or carboxy-terminal Myc-tagged Orai1 in SCID cells using a bicistronic IRES–GFP retroviral vector (Fig. 4d). Non-permeabilized cells did not stain with an anti-Myc antibody (unpublished data), consistent with a topology in which both the N

and C termini are cytoplasmically oriented and so inaccessible to the antibody (Fig. 4c).

Orai1 restores store-operated Ca^{2+} influx in the SCID T cells

Expression of wild-type Orai1 (Orai1(WT)) complements the Ca^{2+} influx defect in SCID T cells and fibroblasts, but mutant R91W Orai1 (Myc–Orai1(R91W)) does not (Fig. 5a–d and Supplementary Fig. 3). Myc-tagged Orai1(WT) expressed in SCID cells using an IRES–GFP retroviral vector restored thapsigargin-induced Ca^{2+} influx in GFP-positive but not GFP-negative cells (Fig. 5 and Supplementary Fig. 3). Notably, Ca^{2+} influx was observed only after store depletion with thapsigargin even when $[\text{Ca}^{2+}]_o$ was increased to 20 mM (Supplementary Fig. 3g). This behaviour is a defining feature of store-operated Ca^{2+} entry through CRAC channels, and shows that Orai1 does not encode a constitutively open Ca^{2+} channel.

Mutant and wild-type Orai1 were expressed at equivalent levels and showed similar localization at and near the plasma membrane, as judged by immunocytochemistry (Fig. 4d and Supplementary Fig. 4) and western blotting (data not shown). The pharmacological

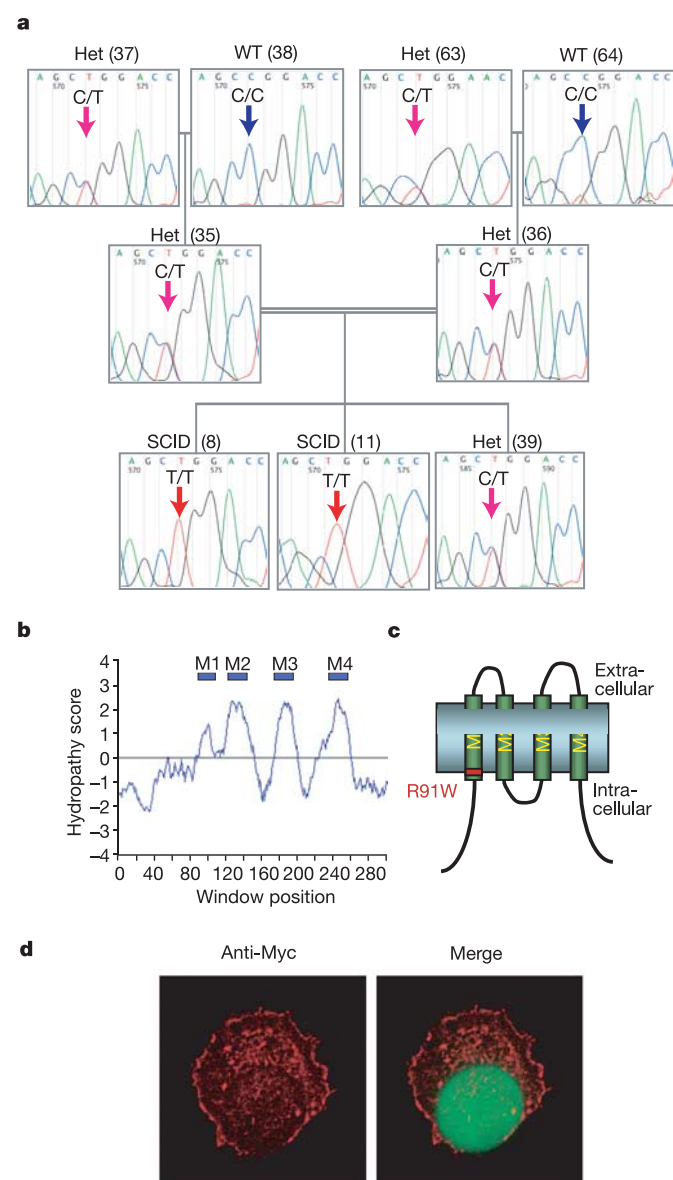


Figure 4 | ORAI1 is mutated in the SCID patients. **a**, Homozygous C→T missense mutation in both SCID patients (8, 11). Other family members (parents, grandparents and unaffected brother) are heterozygous (Het) or wild type (WT) at this position. Identification numbers of individuals as in Fig. 1a. **b**, Hydropathy plot of Orai1 calculated using the Kyte–Doolittle algorithm, with a window size of 19 amino acids. Three transmembrane segments (M2–M4) are predicted to have a score >1.8; M1 has a lower score of ~1.3. **c**, Schematic representation of the predicted membrane topology of Orai1. **d**, Myc-tagged Orai1(WT) (red), coexpressed with GFP (green) in SCID T cells, localizes at or near the plasma membrane.

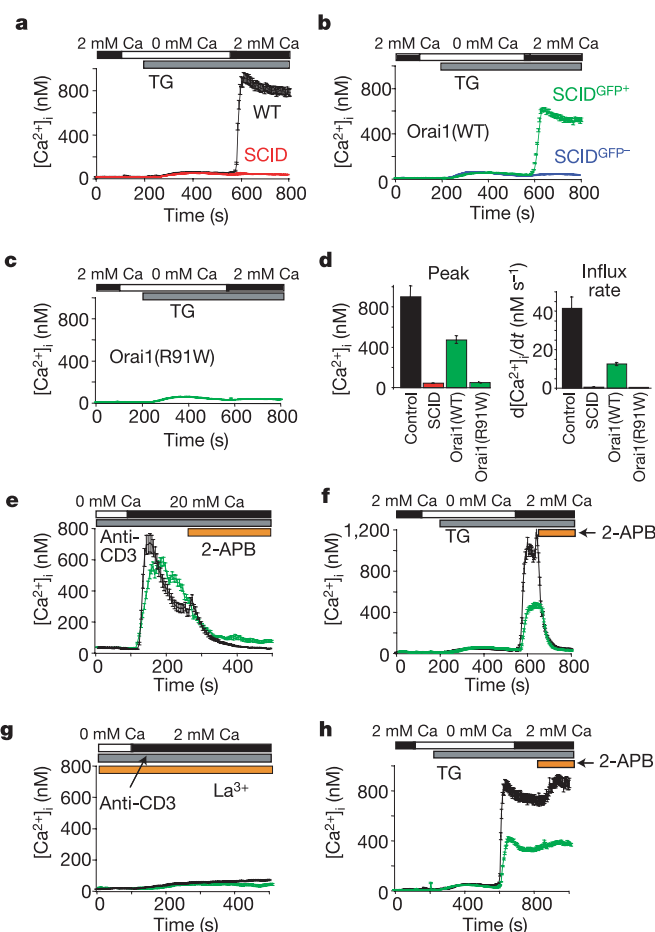


Figure 5 | Orai1 complements the Ca^{2+} entry defect in SCID patients' T cells. **a**, Ca^{2+} influx in untransduced T cells from a control (black) and a SCID patient (red). **b**, **c**, Expression of wild-type Myc–Orai1(WT) (green trace in **b**), but not mutant Myc–Orai1(R91W) (green trace in **c**), in SCID T cells restores store-operated Ca^{2+} entry. **d**, Summary of peak $[\text{Ca}^{2+}]_i$ levels and initial rates of $[\text{Ca}^{2+}]_i$ increase compiled from several experiments similar to those shown in **a–c**. Averages were calculated from all GFP⁺ SCID T cells expressing Orai1(WT) and Orai1(R91W). Error bars represent s.e.m. from three independent experiments (50–100 cells analysed per experiment). **e–h**, Ca^{2+} influx in control T cells (black) and Myc–Orai1(WT)-expressing SCID T cells (green) is inhibited by 75 μM 2-APB (**e**, **f**) or 2 μM La^{3+} (**g**), and enhanced by 3 μM 2-APB (**h**) after stimulation with an anti-CD3 antibody (**e**, **g**) or thapsigargin (**f**, **h**).

characteristics of store-operated Ca^{2+} entry in Orai1-complemented cells were exactly those expected for Ca^{2+} influx through CRAC channels^{7,35}. Treatment with 75 μM 2-aminoethyl-diphenyl borate (2-APB, a compound known to modulate CRAC channel activity) or 2 μM La^{3+} inhibited Ca^{2+} influx, whereas treatment with a low dose of 2-APB (3 μM) caused a distinct further increase in $[\text{Ca}^{2+}]_i$ (Fig. 5e–h and Supplementary Fig. 3). Together, these results show that *ORAI1* is the gene responsible for the Ca^{2+} influx defect in the T cells and fibroblasts of SCID patients.

Expression of Orai1 restores I_{CRAC} in the SCID T cells

Using a whole-cell patch-clamp configuration, we showed that the currents arising from store-depletion in reconstituted SCID T cells have many key features of I_{CRAC} (Fig. 6 and Supplementary Fig. 5).

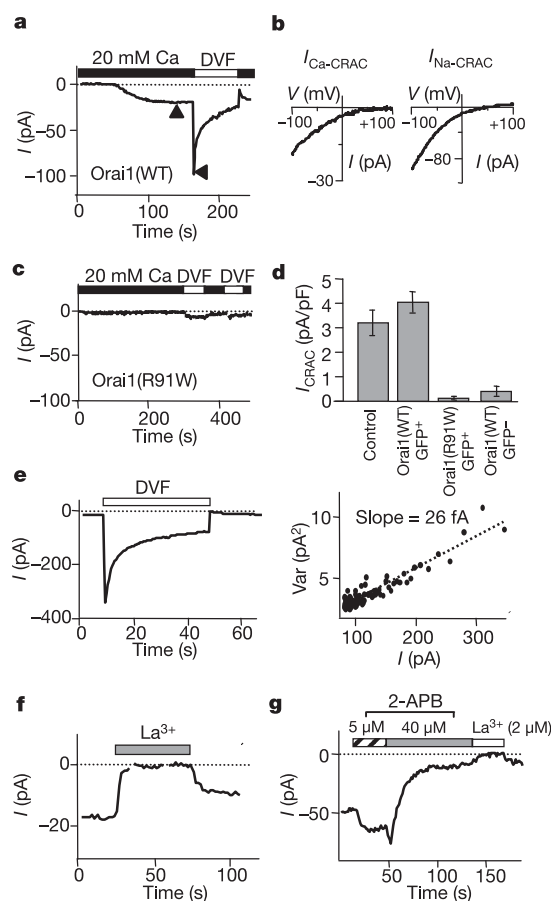


Figure 6 | Expression of Orai1 restores CRAC channel function in SCID T cells. **a**, Ca^{2+} and Na^{+} CRAC currents in an Orai(WT)-complemented SCID T cell after passive store depletion with a pipette solution containing 8 mM BAPTA. **b**, Current–voltage (I – V) relationship in 20 mM $[\text{Ca}^{2+}]_o$ (left) and in divalent-free solution (DVF, right), collected at the times indicated by the arrows in **a**. **c**, In SCID T cells expressing mutant Orai1(R91W), Ca^{2+} and Na^{+} currents fail to develop during passive store depletion by 8 mM BAPTA. **d**, Summary of peak current densities in the indicated cell categories. Cell numbers: control ($n = 5$), Orai1(WT) GFP⁺ ($n = 10$), Orai1(R91W) GFP⁺ ($n = 5$), Orai1(WT) GFP[−] ($n = 4$). Error bars represent s.e.m. **e**, Noise analysis of the depotentiating Na^{+} current measured at a constant potential of -100 mV. Left plot shows the mean current; right plot shows the variance (Var) plotted against the mean current. **f**, Blockade of the Ca^{2+} current by 2 μM La^{3+} . **g**, Potentiation and blockade of I_{CRAC} by low (5 μM) and high (40 μM) doses of 2-APB. Results are representative of 5 out of 5 cells. Panels **a**, **c**, **f** and **g** show peak currents elicited by hyperpolarizing steps to -100 mV. The leak-corrected zero current level is indicated by a dotted line (**a**, **c**, **e**, **f**, **g**). Cells expressing high GFP and Orai1(WT) or Orai1(R91W) were selected for patch-clamp recordings.

In SCID T cells reconstituted with wild-type but not mutant Myc–Orai1, inclusion of 8 mM BAPTA (a Ca^{2+} chelator) in the patch pipette caused slow development of an inward current in 20 mM $[\text{Ca}^{2+}]_o$ reminiscent of the development of I_{CRAC} in response to store depletion with BAPTA or thapsigargin^{5,6} (Fig. 6a). In divalent-free solution lacking Ca^{2+} and Mg^{2+} , an inward Na^{+} current was observed that was initially much larger than the Ca^{2+} current but subsequently declined over tens of seconds (Fig. 6a). This phenomenon, known as depotentiation, is characteristic of CRAC channels in Jurkat T cells, rat basophilic leukaemia (RBL) cells and human T-cell lines^{9,16,36}.

Both the Ca^{2+} and Na^{+} currents showed an inwardly rectifying current–voltage (I – V) relationship (Fig. 6b). The reversal potential of the inward current in 20 mM Ca^{2+} was greater than $+90$ mV, consistent with the known high selectivity of CRAC channels for Ca^{2+} . The reversal potential in divalent-free solution was 49 ± 2 mV ($n = 4$ cells), indicating that, as expected^{16,37}, the channels are only weakly permeable to the Cs^{+} ions in the patch pipette ($P_{\text{Cs}}/P_{\text{Na}} = 0.14$). The noise characteristics of the current were consistent with those of CRAC channels (Fig. 6e): during depotentiation of the Na^{+} current, variance declined linearly with mean current, with an average slope of 29 ± 4 fA ($n = 4$ cells), providing a lower boundary for the unitary current similar to that of previous measurements of I_{CRAC} ¹⁶. The Ca^{2+} current showed fast inactivation in 20 mM $[\text{Ca}^{2+}]_o$, and the extent and time course of inactivation was similar to that previously reported for CRAC channels in Jurkat T cells³⁸ (Supplementary Fig. 5d). The pharmacological hallmarks of the reconstituted current included complete block by 2 μM La^{3+} (Fig. 6f), as well as potentiation by low doses and inhibition by high doses of 2-APB (Fig. 6g). Moreover, the block observed with high doses of 2-APB was accompanied by the loss of fast inactivation³⁵ (data not shown). Together, these studies show that a single point mutation in *ORAI1* disrupts CRAC channel function in T cells.

Discussion

We have used two independent genetic analyses to identify Orai1 as an evolutionarily conserved and essential component of the store-operated Ca^{2+} entry mechanism. Genome-wide SNP analysis of SCID patients and their relatives, analysed using a powerful combination of recessive and dominant linkage mapping, pinpointed a genomic region with a very high probability of linkage to the mutant gene. A parallel genome-wide screen in *Drosophila* identified several candidates for which RNAi-mediated depletion interfered with thapsigargin-mediated nuclear localization of NFAT–GFP. These candidates included *dStim*, a known ER Ca^{2+} sensor and regulator of store-operated Ca^{2+} entry^{24–26}, and a novel candidate, *olf186-F* (here renamed *Drosophila Orai*). Validating our dual strategy, the gene encoding one of the human homologues of dOrai, Orai1 (hypothetical protein FLJ14466), is located on chromosome 12q24, exactly the region identified through our SNP analysis as genetically linked to the SCID syndrome. Additional studies confirmed that a point mutation in *ORAI1* is responsible for the genetic defect in store-operated Ca^{2+} entry and I_{CRAC} in cells of the SCID patients^{9,13}. It will be interesting to determine whether *ORAI1* is also mutated in two other SCID patients identified as having similar defects in CRAC channel function^{10,39}, or whether these patients have defects in some other component of the signalling pathway leading to CRAC channel activation.

The genetic basis for the SCID defect is not a known polymorphism, nor is it represented in the 270 individuals of the HapMap panel³³, a number sufficient to find almost all haplotypes with frequencies $\geq 5\%$. The possibility that the C→T mutation is a single-nucleotide polymorphism confined to a small ethnic population not represented in the HapMap panel can be ruled out with reasonable certainty, based on the fact that complementation with Orai1 restores store-operated Ca^{2+} entry and I_{CRAC} in SCID patient

cells. Furthermore, arginine 91, which is mutated in the SCID patients, is located in a putative transmembrane region that is conserved across species (Fig. 4b, c and Supplementary Fig. 2), highlighting its potential importance in Orai1 function.

ORAI1 mRNA is broadly expressed in mammalian tissues (data not shown), potentially explaining previous observations that not only the T cells from SCID patients, but also B cells and fibroblasts, have a substantial defect in store-operated Ca^{2+} entry¹³. Surprisingly, however, the clinical phenotype of the SCID patients is predominantly one of immunodeficiency, associated in the single surviving patient with ectodermal dysplasia and anhydrosis (EDA) and a congenital, non-progressive myopathy (unpublished observations). EDA is characterized by defects in the morphogenesis of ectodermal structures, including hair, skin, sweat glands and teeth, and previous studies have linked it to hypoactivation of the transcription factor NF- κ B^{40–43}. Ca^{2+} mobilization is thought to contribute to NF- κ B activation in T cells and other cell types under certain stimulatory conditions⁴⁴, thus the EDA syndrome may reflect defective NF- κ B activation, either during development or acutely in specific cell types. In contrast, the myopathy could potentially be a direct consequence of defective NFAT activation, given that NFAT has a substantial influence on certain aspects of skeletal muscle development and function (reviewed in refs 3, 45).

The characteristics of Ca^{2+} influx and the Ca^{2+} current in Orai1-complemented SCID T cells are indistinguishable from those observed in control T cells. Both are strictly regulated by store depletion, and the electrophysiological and pharmacological properties of the restored current are fully consistent with those of I_{CRAC} . Although the specific role of Orai1 has not yet been determined, the available data are consistent with the possibility that Orai1 encodes a channel subunit or a closely associated channel regulator in the plasma membrane. The hydropathy profile of Orai1 predicts a membrane protein with four transmembrane segments (Fig. 4b, c). Immunocytochemistry of Myc-tagged Orai1 in resting cells is consistent with localization at the plasma membrane (Fig. 4d), and both N- and C-terminal epitope tags on Orai1 are inaccessible to antibody staining in non-permeabilized cells (data not shown), compatible with the topology of a plasma membrane channel subunit in which both N and C termini are cytoplasmic (Fig. 4c). The plasma membrane localization of Orai1 differs from that of STIM1, which is predominantly located in the ER^{9,25,26}. Future studies will address the relation between Orai and Stim proteins and the mechanism by which store depletion couples to CRAC channel opening.

METHODS

The *Drosophila* genome-wide RNAi screen was performed as previously described^{27–29}. Detailed descriptions of all methods are provided in the online Supplementary Information.

Single-nucleotide polymorphism array-based linkage analysis. Genomic DNA of SCID patients and 21 relatives was prepared from peripheral blood mononuclear cells using genomic DNA maxi prep kits (Qiagen). Genotyping was performed at the SNP Genotyping Center of the Broad Institute and the Harvard Partners Center for Genetics and Genomics, using Affymetrix Human Mapping 10K (Xba 142 2.0) microarrays with an average SNP heterozygosity of 0.38 and a mean intermarker distance of 258 kb, allowing for simultaneous genotyping of more than 10,000 SNPs in the human genome.

For parametric linkage analysis, data were converted into 'Linkage' format using 'Compare Linkage'⁴⁶, and imported into the multi-point linkage software programs Merlin and Allegro. Mendelian genotype errors inconsistent with the parental genotypes were detected and set to missing genotypes. Multipoint parametric linkage analysis was performed to compute LOD scores at each SNP position using Allegro⁴⁷, as shown in Fig. 2. Linkage analysis using Genehunter 2.1r6 provided very similar results. For parametric analysis, a disease allele frequency of 0.001, a penetrance value of 0.99 and a phenocopy of 0.01 were used for all of the pedigrees. Two distinct parametric linkage analyses were carried out using recessive and dominant models of inheritance, respectively. For the 'recessive' model, haplotypes from both patients, their parents, unaffected brother and grandparents (individuals 8, 11, 35, 36, 37, 38, 39, 63 and 64 in

Fig. 1a) were analysed assuming an autosomal recessive mode of inheritance for the SCID disease. The predicted maximum LOD score from this analysis was ~ 1.9 (that is, $-\log_{10}[0.25 \times 0.25 \times 0.25 \times 0.75]$). For the 'dominant' model, 12 family members with reduced store-operated Ca^{2+} entry were defined as 'affected' (individuals 31, 32, 34, 35, 36, 37, 41, 51, 53, 56, 57 and 63 in Fig. 1a)—that is, carriers of a dominantly acting mutation—and their SNP haplotypes were compared to those of eight healthy family members with normal store-operated Ca^{2+} entry. The predicted maximum LOD score from this analysis was ~ 3.6 (that is, $-\log_{10}[0.5^{12}]$).

Received 2 February; accepted 7 March 2006.

Published online 2 April 2006; corrected 11 May 2006 (details online).

- Lewis, R. S. Calcium signaling mechanisms in T lymphocytes. *Annu. Rev. Immunol.* **19**, 497–521 (2001).
- Feske, S., Okamura, H., Hogan, P. G. & Rao, A. Ca^{2+} /calcineurin signalling in cells of the immune system. *Biochem. Biophys. Res. Commun.* **311**, 1117–1132 (2003).
- Hogan, P. G., Chen, L., Nardone, J. & Rao, A. Transcriptional regulation by calcium, calcineurin, and NFAT. *Genes Dev.* **17**, 2205–2232 (2003).
- Gallo, E. M., Cante-Barrett, K. & Crabtree, G. R. Lymphocyte calcium signaling from membrane to nucleus. *Nature Immunol.* **7**, 25–32 (2006).
- Hoth, M. & Penner, R. Depletion of intracellular calcium stores activates a calcium current in mast cells. *Nature* **355**, 353–356 (1992).
- Zweifach, A. & Lewis, R. S. Mitogen-regulated Ca^{2+} current of T lymphocytes is activated by depletion of intracellular Ca^{2+} stores. *Proc. Natl Acad. Sci. USA* **90**, 6295–6299 (1993).
- Parekh, A. B. & Putney, J. W. Jr. Store-operated calcium channels. *Physiol. Rev.* **85**, 757–810 (2005).
- Prakriya, M. & Lewis, R. S. CRAC channels: activation, permeation, and the search for a molecular identity. *Cell Calcium* **33**, 311–321 (2003).
- Feske, S., Prakriya, M., Rao, A. & Lewis, R. S. A severe defect in CRAC Ca^{2+} channel activation and altered K^{+} channel gating in T cells from immunodeficient patients. *J. Exp. Med.* **202**, 651–662 (2005).
- Partiseti, M. et al. The calcium current activated by T cell receptor and store depletion in human lymphocytes is absent in a primary immunodeficiency. *J. Biol. Chem.* **269**, 32327–32335 (1994).
- Schlesier, M. et al. Primary severe immunodeficiency due to impaired signal transduction in T cells. *Immunodeficiency* **4**, 133–136 (1993).
- Feske, S. et al. Severe combined immunodeficiency due to defective binding of the nuclear factor of activated T cells in T lymphocytes of two male siblings. *Eur. J. Immunol.* **26**, 2119–2126 (1996).
- Feske, S., Giltner, J., Dolmetsch, R., Staudt, L. & Rao, A. Gene regulation by calcium influx in T lymphocytes. *Nature Immunol.* **2**, 316–324 (2001).
- Feske, S., Draeger, R., Peter, H. H., Eichmann, K. & Rao, A. The duration of nuclear residence of NFAT determines the pattern of cytokine expression in human SCID T cells. *J. Immunol.* **165**, 297–305 (2000).
- Feske, S., Draeger, R., Peter, H. H. & Rao, A. Impaired NFAT regulation and its role in a severe combined immunodeficiency. *Immunobiology* **202**, 134–150 (2000).
- Prakriya, M. & Lewis, R. S. Separation and characterization of currents through store-operated CRAC channels and Mg^{2+} -inhibited cation (MIC) channels. *J. Gen. Physiol.* **119**, 487–507 (2002).
- Mori, Y. et al. Transient receptor potential 1 regulates capacitative Ca^{2+} entry and Ca^{2+} release from endoplasmic reticulum in B lymphocytes. *J. Exp. Med.* **195**, 673–681 (2002).
- Philipp, S. et al. TRPC3 mediates T-cell receptor-dependent calcium entry in human T-lymphocytes. *J. Biol. Chem.* **278**, 26629–26638 (2003).
- Cui, J., Bian, J. S., Kagan, A. & McDonald, T. V. CaT1 contributes to the store-operated calcium current in Jurkat T-lymphocytes. *J. Biol. Chem.* **277**, 47175–47183 (2002).
- Yue, L., Peng, J. B., Hediger, M. A. & Clapham, D. E. CaT1 manifests the pore properties of the calcium-release-activated calcium channel. *Nature* **410**, 705–709 (2001).
- Voets, T. et al. CaT1 and the calcium release-activated calcium channel manifest distinct pore properties. *J. Biol. Chem.* **276**, 47767–47770 (2001).
- Clapham, D. E. TRP channels as cellular sensors. *Nature* **426**, 517–524 (2003).
- Putney, J. W. Jr. Capacitative calcium entry: sensing the calcium stores. *J. Cell Biol.* **169**, 381–382 (2005).
- Roos, J. et al. STIM1, an essential and conserved component of store-operated Ca^{2+} channel function. *J. Cell Biol.* **169**, 435–445 (2005).
- Zhang, S. L. et al. STIM1 is a Ca^{2+} sensor that activates CRAC channels and migrates from the Ca^{2+} store to the plasma membrane. *Nature* **437**, 902–905 (2005).
- Liou, J. et al. STIM is a Ca^{2+} sensor essential for Ca^{2+} -store-depletion-triggered Ca^{2+} influx. *Curr. Biol.* **15**, 1235–1241 (2005).
- Armknicht, S. et al. High-throughput RNA interference screens in *Drosophila* tissue culture cells. *Methods Enzymol.* **392**, 55–73 (2005).
- Boutros, M. et al. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832–835 (2004).

29. Gwack, Y. *et al.* A genome-wide *Drosophila* RNAi screen identifies DYRK as a novel regulator of NFAT. *Nature* advance online publication, doi:10.1038/nature04631 (1 March 2006).
30. Yeromin, A. V., Roos, J., Stauderman, K. A. & Cahalan, M. D. A store-operated calcium channel in *Drosophila* S2 cells. *J. Gen. Physiol.* **123**, 167–182 (2004).
31. Homer. *The Iliad* (Book 5, line 859; Book 8, line 449) (Viking, New York, 1990).
32. Stewart, M. "The Hours", *Greek Mythology: From the Iliad to the Fall of the Last Tyrant*. <http://messagenet.com/myths/bios/hours.html> (2005).
33. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
34. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
35. Prakriya, M. & Lewis, R. S. Potentiation and inhibition of Ca^{2+} release-activated Ca^{2+} channels by 2-aminoethylphenyl borate (2-APB) occurs independently of IP_3 receptors. *J. Physiol. (Lond.)* **536**, 3–19 (2001).
36. Hermosura, M. C., Montell-Zoller, M. K., Scharenberg, A. M., Penner, R. & Fleig, A. Dissociation of the store-operated calcium current I_{CRAC} and the Mg-nucleotide-regulated metal ion current MagNum. *J. Physiol. (Lond.)* **539**, 445–458 (2002).
37. Lippman-Wienhues, A. & Cahalan, M. D. Conductance and permeation of monovalent cations through depletion-activated Ca^{2+} channels I_{CRAC} in Jurkat T cells. *Biophys. J.* **71**, 787–794 (1996).
38. Zweifach, A. & Lewis, R. S. Rapid inactivation of depletion-activated calcium current I_{CRAC} due to local calcium feedback. *J. Gen. Physiol.* **105**, 209–226 (1995).
39. Le Deist, F. *et al.* A primary T-cell immunodeficiency associated with defective transmembrane calcium influx. *Blood* **85**, 1053–1062 (1995).
40. Doffinger, R. *et al.* X-linked anhidrotic ectodermal dysplasia with immunodeficiency is caused by impaired NF- κ B signaling. *Nature Genet.* **27**, 277–285 (2001).
41. Courtis, G. *et al.* A hypermorphic $\text{I}\kappa\text{B}\alpha$ mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. *J. Clin. Invest.* **112**, 1108–1115 (2003).
42. Schmidt-Ullrich, R. *et al.* Requirement of NF- κ B/Rel for the development of hair follicles and other epidermal appendages. *Development* **128**, 3843–3853 (2001).
43. Smahi, A. *et al.* The NF- κ B signalling pathway in human diseases: from incontinentia pigmenti to ectodermal dysplasias and immune-deficiency syndromes. *Hum. Mol. Genet.* **11**, 2371–2375 (2002).
44. Kanno, T. & Siebenlist, U. Activation of nuclear factor- κ B via T cell receptor requires a Raf kinase and Ca^{2+} influx. Functional synergy between Raf and calcineurin. *J. Immunol.* **157**, 5277–5283 (1996).
45. Crabtree, G. R. & Olson, E. N. NFAT signaling: Choreographing the social lives of cells. *Cell* **109**, S67–S79 (2002).
46. Leykin, I. *et al.* Comparative linkage analysis and visualization of high-density oligonucleotide SNP array data. *BMC Genet.* **6**, 7 (2005).
47. Gudbjartsson, D. F., Jonasson, K., Frigge, M. L. & Kong, A. Allegro, a new computer program for multipoint linkage analysis. *Nature Genet.* **25**, 12–13 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Ehl, I. Barlan, G. Tuncman and T. Akkoc for facilitating the contact with the SCID patients' families; B. Mathey-Prevot, N. Perrimon, and staff at the *Drosophila* RNAi Screening Centre at Harvard Medical School for valuable assistance with the screens; M. Komarinski for software support (GeneHunter 2.1r5); J. Nardone for help with sequence analysis; S. Sharma for providing the S2R + cell line stably transduced with NFAT-GFP; and D. Neems for technical assistance. This work was funded by grants from the National Institutes of Health to A.R., R.S.L. and S.F., and an Immune Deficiency Foundation grant to S.F.

Author Contributions S.F. was responsible for all experiments involving genetic and functional analysis of SCID patients, and was assisted in these experiments by S.-H.P. Y.G. designed and implemented the genome-wide *Drosophila* RNA screen that identified dOrai, with assistance from S.S. M.P. and R.S.L. analysed the electrophysiological properties of calcium currents in Orai1-reconstituted SCID T cells provided by S.F. M.D. advised on the design of the linkage mapping screen and analysis of linkage data, which was carried out by B.T. P.H. and A.R. provided advice and overall direction, and supervised project planning and execution.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.R. (arao@cbr.med.harvard.edu).

ARTICLES

Seawater subduction controls the heavy noble gas composition of the mantle

Greg Holland¹ & Chris J. Ballentine¹

The relationship between solar volatiles and those now in the Earth's atmosphere and mantle reservoirs provides insight into the processes controlling the acquisition of volatiles during planetary accretion and their subsequent evolution. Whereas the light noble gases (helium and neon) in the Earth's mantle preserve a solar-like isotopic composition, heavy noble gases (argon, krypton and xenon) have an isotopic composition very similar to that of the modern atmosphere, with radiogenic and (in the case of xenon) solar contributions. Mantle noble gases in a magmatic CO₂ natural gas field have been previously corrected for shallow atmosphere/groundwater and crustal additions. Here we analyse new data from this field and show that the elemental composition of non-radiogenic heavy noble gases in the mantle is remarkably similar to that of sea water. We challenge the popular concept of a noble gas 'subduction barrier'—the convecting mantle noble gas isotopic and elemental composition is explained by subduction of sediment and seawater-dominated pore fluids. This accounts for ~100% of the non-radiogenic argon and krypton and 80% of the xenon. Approximately 50% of the convecting mantle water concentration can then be explained by this mechanism. Enhanced recycling of subducted material to the mantle plume source region then accounts for the lower ratio of radiogenic to non-radiogenic heavy noble gas isotopes and higher water content of plume-derived basalts.

Owing to their low concentrations and unreactive nature, noble gases are key tracers of both mantle evolution and the origin of related major volatiles such as H₂O and CO₂. The present day complement of noble gases in the terrestrial mantle is a function of several processes. Observed He, Ne and Xe isotopes^{1–3} in the mantle require that the Earth trapped solar-like volatiles during the accretionary process ~4.55 Gyr ago. Superimposed on this primordial signature are radiogenic/nucleogenic noble gases as well as heavy noble gases (Ar, Kr and Xe) isotopically identical to the atmosphere. Nevertheless, early models of mantle noble gas behaviour explicitly require a subduction barrier for noble gases⁶, while others attribute volatile-rich ocean island basalts (OIBs) to assimilation of sea water during eruption rather than as an integral part of the mantle source^{7,8}. Also, no current model can account for the modified noble gas isotopic composition of the terrestrial atmosphere without early outgassing from a mantle with a solar-like isotopic composition⁹. Thus far, however, no reproducible deviations from air values in non-radiogenic Ar and Kr isotopes in mantle samples have been observed^{10,11}. In mantle steady-state models^{12,13}, air recycling of heavy noble gases into the convecting mantle is a parameter that depends on the composition of an isolated deep mantle.

The issue is compounded by the ubiquitous problem of local air contamination in all terrestrial basalt samples^{14,15}, which occurs by crustal assimilation or during/after eruption on exposure to sea water/atmosphere. The lack of information regarding the extent of recycling and the mass balance between accretionary and recycled volatiles hampers interpretation of data from plume-related 'hot-spots' with high ³He/⁴He ratios and the convecting mantle sampled at mid-ocean ridges. Previous work from CO₂ well gases has now identified the character and origin of the light noble gases in the convecting mantle⁴. In this work, we extend these high precision analyses to Xe, which, critically, is alone among the heavy noble gases in showing a non-air isotopic composition in the non-radiogenic

isotopes. The nature of the sample suite uniquely permits unambiguous removal of all local air-derived (groundwater) addition, and now allows us to quantitatively address the issue of volatile recycling into the mantle.

Xe isotopic composition of the Bravo Dome gas field

The Bravo Dome CO₂ gas field in New Mexico, USA, was the focus of a recent mantle noble gas study⁴ and the detailed geology and background references are given there. The samples from this earlier study were unavailable for further analysis and new samples were collected from producing wells in this field in September 2004, with emphasis on collection from the western section of the gas field, which is rich in mantle-derived noble gases. One sample, WBD-4, was from a well that had been non-producing for 7 days. ³He/⁴He, ²⁰Ne/²¹Ne/²²Ne, ³⁶Ar/³⁸Ar/⁴⁰Ar, ⁸⁴Kr and Xe abundance and isotopic composition were determined in the Manchester University noble gas laboratory (Tables 1–3). In addition, 15 samples were collected from Sheep Mountain, Colorado, and 5 samples were collected from McElmo Dome, Colorado. Xe isotopic data only are presented for these two fields (see Supplementary Information).

Air and crustal-radiogenic Ne in the gas field are mixed in almost constant proportion, before variably mixing with the magmatic CO₂ (ref. 4). This generates an (air+crust)–mantle mixing line, which intersects with the air–mid-ocean-ridge basalt (MORB) mixing line to uniquely define the mantle Ne isotopic composition. The Xe isotope ratios of Bravo Dome samples show the same systematics, with a pre-mixed (air+crust) Xe component variably mixing with mantle Xe. The intersection of the (air+crust)–mantle Xe mixing line with the air–MORB Xe mixing line enables direct determination of the ¹²⁹Xe excess in the convecting mantle for the first time, where ¹²⁹Xe/¹³⁰Xe_(mantle) = 7.90 ± 0.14 (Fig. 1). Excesses in ^{124–126–128}Xe/¹³⁰Xe relative to air are similar to the single sample measured previously⁵, although with all shielded isotopes plotting on

¹School of Earth, Atmosphere and Environmental Sciences, The University of Manchester, Oxford Road, Manchester M13 9PL, UK.

Table 1 | Noble gas elemental data

Sample	^4He (10^{-5})	Error (10^{-6})	^{20}Ne (10^{-9})	Error (10^{-11})	^{40}Ar (10^{-5})	Error (10^{-7})	^{84}Kr (10^{-11})	Error (10^{-12})	^{132}Xe (10^{-11})	Error (10^{-13})
BD11-B	3.93	1.1	1.20	2.89	2.79	3.81	3.46	8.06	0.669	5.72
BD20-B	4.17	5.0	1.21	2.92	2.95	4.00	3.68	5.42	0.670	2.81
WBD04-B	4.01	4.0	1.09	2.59	3.07	4.16	4.23	2.51	0.687	2.79
BD19-B	9.87	1.1	2.43	5.76	4.33	5.86	15.1	10.0	1.78	9.27
BD15-B	5.83	3.3	1.70	4.03	3.11	4.21	5.14	9.73	1.04	4.41
BD02-B	47.7	1.4	9.43	2.23	7.72	10.5	62.7	10.4	6.56	26.7
BD07-B	8.22	0.95	1.93	4.64	3.48	4.71	10.5	7.09	1.31	6.64
BD13-B	15.4	1.9	2.97	7.04	4.50	6.09	19.9	4.75	2.12	8.89
BD05-B	28.0	1.2	5.08	1.20	6.00	8.13	3.53	14.6	4.10	17.0
BD04-B	9.84	2.4	2.35	5.65	3.31	4.48	11.0	8.32	1.40	6.05
WBD01-B	5.58	0.47	1.44	3.43	3.17	4.28	5.42	4.33	0.825	3.40
WBD03-B	6.64	0.50	1.60	3.80	3.11	4.21	5.52	8.36	0.922	3.39

All values are corrected for full procedural blanks. Concentrations are in cm^3 STP per cm^3 . All quoted errors are 1σ . Sample nomenclature: prefix BD, Bravo Dome; WBD, West Bravo Dome.

the air–solar mixing line (Fig. 2). The least air-contaminated sample requires a $10 \pm 4\%$ contribution by a solar Xe component. From Fig. 1, it can be clearly seen that even the least air-contaminated samples still contain $\sim 50\%$ (air+crust) Xe. Correcting for this local air+crust addition allows us to now define the proportion of solar Xe ($20 \pm 7\%$) to Xe with an air-like ratio (80%) in the convecting mantle.

In a three-dimensional plot of $i^{22}\text{Ne}$ versus $^{21}\text{Ne}/^{22}\text{Ne}$ and $^{20}\text{Ne}/^{22}\text{Ne}$, where i is any noble gas isotope, a data set that is a mixture of the same three end-member components will fall on a plane bounded by the crustal-radiogenic, mantle and air end-members. This is shown for $i = ^{130}\text{Xe}$ in Fig. 3. If the mantle Ne isotopic composition is known, this enables the mantle $i^{22}\text{Ne}$ composition for this sample suite to be determined⁴. Using the new Ne isotope data determined in this study, with more samples from the mantle-rich section of the field, we are able to refine the Ne isotopic estimate of the mantle end-member to be $^{20}\text{Ne}/^{22}\text{Ne} = 12.49 \pm 0.04$ and $^{21}\text{Ne}/^{22}\text{Ne} = 0.0578 \pm 0.0003$ (Supplementary Information). We use these new values to calculate the Bravo Dome mantle $i^{22}\text{Ne}$ composition, where $i = ^{36}\text{Ar}$, ^{84}Kr and ^{130}Xe . In comparing these data to other Solar System reservoirs, we re-normalize to ^{36}Ar and plot $(i^{36}\text{Ar})_{\text{sample}}/(i^{36}\text{Ar})_{\text{solar}}$ (Fig. 4). Measured elemental and isotopic ratios are also presented in Tables 1–3.

Evidence for seawater recycling of noble gases

Deriving convecting mantle elemental ratios from MORB and OIB measurements is complicated by eruption-related fractionation. The issue is further confused, because in addition to ubiquitous unfractionated modern-air contamination¹⁴, elementally fractionated components with an air-like isotopic composition are also observed¹⁵. Nevertheless, eruption-related fractionation is probably due to partitioning between vesicles and melt and subsequent vesicle loss at elevated pressure¹⁶. In this context, highly vesicular MORB glasses

may represent the least elementally fractionated basalt samples¹⁷. The highly vesicular 2IID43 reference ‘popping rock’ has minimal air contamination, evidenced by preserving noble gas isotopic compositions indistinguishable from the upper mantle values defined by the well gas data set^{4,18,19} (Table 4). We observe that the mantle supplying the Bravo Dome has an elemental composition very similar to the least-contaminated gases released from the 2IID43 popping rock (Fig. 4). It is highly unlikely that two mantle volatile systems, formed in very different ways, would be elementally fractionated to the same extent. We argue that the similarity in elemental and isotopic composition of the two systems is strong *prima-facie* evidence that both the 2IID43 popping rock and the mantle source supplying the Bravo Dome well gases are elementally unfractionated samples from the same mantle source.

We can now show that the convecting mantle $^{84}\text{Kr}/^{36}\text{Ar}$ ratio is distinct from modern air and that the $^{130}\text{Xe}/^{36}\text{Ar}$ ratio is an order of magnitude higher than air (Fig. 4). In contrast, the elemental abundance pattern of the heavy noble gases is remarkably close to that of sea water. Given that subduction zones lie under several kilometres of sea water this is, perhaps, no surprise. In detail, the convecting mantle $^{84}\text{Kr}/^{36}\text{Ar}$ ratio is slightly enriched in Kr relative to sea water, and the $^{130}\text{Xe}/^{36}\text{Ar}$ ratio is within a factor of 2 of sea water (Fig. 4) after correcting the convecting mantle for the $\sim 20\%$ solar Xe. This relatively small excess of Xe in the convecting mantle can be readily attributed to addition of small amounts of oceanic sediment that is enriched in the heavy noble gases relative to air and sea water²⁰. We calculate that the amount of sedimentary Xe required to explain elevated $^{36}\text{Ar}/^{130}\text{Xe}$ will generate $^{84}\text{Kr}/^{36}\text{Ar} = 0.053$. This is indistinguishable from the observed ratio, and accounts for the small relative Kr excess relative to seawater values. We argue that seawater recycling into the mantle with a small sedimentary component accounts for both the elemental and isotopic composition of the non-radiogenic heavy noble gases.

The ratios $^{84}\text{Kr}/^{36}\text{Ar}$ and $^{130}\text{Xe}/^{36}\text{Ar}$ in sea water are dominantly

Table 2 | Isotopic and component ratios

Sample	$^3\text{He}/^4\text{He}$	Error	$^{20}\text{Ne}/^{22}\text{Ne}$	Error	$^{21}\text{Ne}/^{22}\text{Ne}$	Error	$^{40}\text{Ar}/^{36}\text{Ar}$	Error	$^{38}\text{Ar}/^{36}\text{Ar}$	Error
BD11-B	3.681	0.059	11.65	0.14	0.0568	0.0007	19,733	300	0.1878	0.0004
BD20-B	3.720	0.067	11.85	0.04	0.0571	0.0003	21,401	392	0.1886	0.0009
WBD04-B	3.839	0.065	11.79	0.04	0.0568	0.0003	22,548	730	0.1884	0.0008
BD19-B	1.888	0.039	10.70	0.04	0.0543	0.0002	9,888	152	0.1887	0.0005
BD15-B	2.424	0.043	10.88	0.07	0.0514	0.0004	13,472	224	0.1886	0.0005
BD02-B	0.711	0.062	9.84	0.03	0.0500	0.0002	4,197	69	0.1880	0.0004
BD07-B	2.092	0.070	10.89	0.09	0.0554	0.0005	11,205	172	0.1890	0.0004
BD13-B	1.275	0.066	10.45	0.05	0.0582	0.0003	7,765	117	0.1878	0.0004
BD05-B	0.884	0.016	10.04	0.03	0.0541	0.0002	5,288	81	0.1883	0.0004
BD04-B	1.513	0.028	10.45	0.05	0.0517	0.0003	9,612	133	0.1884	0.0006
WBD01-B	3.154	0.061	11.47	0.06	0.0565	0.0003	18,496	298	0.1879	0.0005
WBD03-B	2.526	0.071	11.23	0.03	0.0556	0.0002	16,792	268	0.1881	0.0010

Errors quoted on all ratios are 1σ . $^3\text{He}/^4\text{He}$ are displayed relative to the atmospheric ratio R_a (that is, as R/R_a); $R_a = 1.4 \times 10^{-6}$.

controlled by the relative partial pressure of Ar/Kr/Xe in the atmosphere and to a minor extent by the temperature and salinity of the oceans. We make the *a priori* assumption that the atmosphere is not related to other reservoirs and serves only as a source for subducted noble gases, the relative concentrations of which have remained unchanged since atmosphere formation⁹. Seawater $^{84}\text{Kr}/^{36}\text{Ar}$ and $^{130}\text{Xe}/^{36}\text{Ar}$ have therefore remained nearly constant over the lifetime of the Earth. Although subduction occurs in a seawater-dominated fluid environment, the noble gas abundance pattern in any water phase is highly sensitive to physical processes of fractionation. Because the relative solubilities of Ar and Xe can be an order of magnitude different, even a small amount of phase separation will result in significant elemental fractionation. The absence of fractionation from the seawater heavy noble gas abundance pattern suggests that the noble gases are unlikely to have been decoupled from the host sea water. It is reasonable to assume that if recycling of sea water is responsible for the observed heavy noble gas pattern of the convecting mantle, there will be a proportional volume of sea water subducted with the atmosphere-derived heavy noble gases. From Fig. 4, it is clear that seawater recycling of He and Ne into the mantle has negligible effect on the mantle composition owing to the low concentration of He and Ne in sea water, and He/Ne recycling is not discussed further. We further note that contributions from non-seawater/sediment sources to either Ar or Kr are thus negligible.

Seawater-derived water in the convecting mantle

Recycling sea water back into the convecting mantle means that the mantle noble gas concentrations then place a minimum constraint on the recycled water content of the convecting (MORB-source)

mantle. Using the present day ^3He flux into the ocean, the average MORB generation rate and assuming 10% partial melting gives a lower limit of 1.2×10^9 atoms $^3\text{He g}^{-1}$ in the convecting mantle. From the 2IID43 popping rock $^3\text{He}/^{36}\text{Ar}$ ratio (Table 4), the ^{36}Ar concentration in the convecting mantle is 1.5×10^9 atoms g^{-1} . Deep sea water contains 3.37×10^{13} atoms of ^{36}Ar per g. Therefore, 44 p.p.m. H_2O in the convecting mantle is associated with recycled ^{36}Ar , and is seawater-derived. If we assume that water behaves incompatibly during melting, then unaltered MORB glasses give a convecting mantle source with 54–120 p.p.m. water^{21,22}. This can be tested for compatibility with the mantle D/H systematics. By definition, the terrestrial oceans have a bulk $\delta\text{D} = 0\text{‰}$ (all δD values are given here with respect to Standard Mean Ocean Water, SMOW). In contrast, the D/H ratio of mid-ocean-ridge samples and therefore the convecting mantle is distinctly lower, with $\delta\text{D} = -71\text{‰}$ to -91‰ . With the simplifying approximation that 50% of the mantle water is seawater-derived, we require a primordial component with $\delta\text{D} \approx -160\text{‰}$. This is compatible with the range of δD (-165‰ to $+90\text{‰}$) spanned by the majority of chondrite bulk analyses²³.

Unambiguous noble gas evidence for recycled sea water in MORB is hard to obtain, with evidence for volatile recycling at subduction zones^{24–27} also ascribed to processes of shallow level or eruption/sampling-related contamination^{14,28}. However, our new data now enable an assessment of the efficiency of seawater recycling to the convecting mantle. Estimates for the ratio of fluxes of water into and out of arc settings suggest the subduction flux associated with hydrated minerals and bound to sediment is a factor of 10 lower

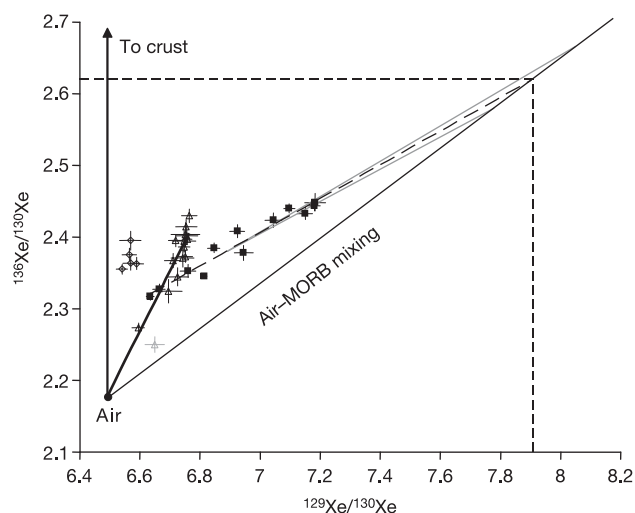


Figure 1 | The intersection of the Bravo Dome Xe array with the MORB (mid-ocean-ridge basalt) array defines the Xe mantle composition. In an analogous way to Ne isotope systematics (Supplementary Information and ref. 4), Xe data from the Bravo Dome system (filled squares) show a two-component air+crust mixture that intersects with the MORB air–mantle mixing line to define the mantle Xe isotopic end-member to be $^{129}\text{Xe}/^{130}\text{Xe}_{(\text{mantle})} = 7.90 \pm 0.14$. The MORB–air mixing line is defined by popping rock data^{19,38}. In contrast, Sheep Mountain data (open triangles) exhibit a mantle–crust mixture that subsequently mixes with air. The grey open triangle is from a well head that had been non-producing for 7 days and is excluded from the fit to the data. McElmo Dome data are represented by open diamonds. Sheep Mountain and McElmo Dome Xe data are presented in Supplementary Information. Vertical and horizontal dashed lines are the intersection of the MORB array with the Bravo Dome array and indicate the convecting mantle values. The dashed line through Bravo Dome data is the error weighted best fit and the dark grey solid lines represent the 1σ error envelope. All errors are 1σ . Line fitting is by the procedure of ref. 48.

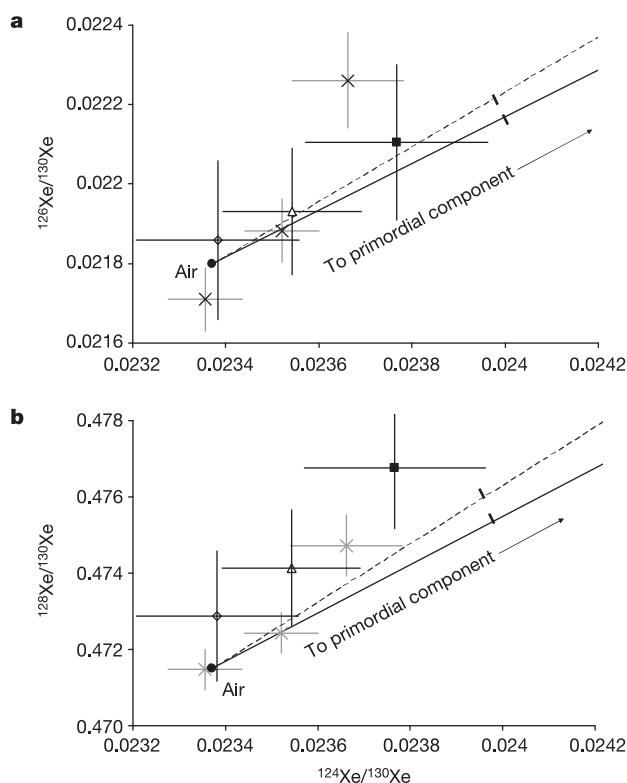


Figure 2 | Non-radiogenic Xe data reveal contribution from a primordial component. **a**, $^{126}\text{Xe}/^{130}\text{Xe}$ versus $^{124}\text{Xe}/^{130}\text{Xe}$; **b**, $^{128}\text{Xe}/^{130}\text{Xe}$ versus $^{124}\text{Xe}/^{130}\text{Xe}$. Plotted data are the least air-contaminated sample from each field as defined by $^{40}\text{Ar}/^{36}\text{Ar}$. Error bars show measurement uncertainty (1σ). Symbols as for Fig. 1. Data from a single gas well⁵ at Bravo Dome, Sheep Mountain and McElmo Dome are included for comparison (cross symbols). Mixing lines between air and a primordial Xe component are also shown: the solid line is air–solar mixing and the dashed line is air–Q mixing, where Q is a planetary Xe component observed in meteorites. Tick marks denote 10% primordial contribution. Primordial component data in refs 9, 49.

Table 3 | Xe isotope data

Sample	¹²⁴ Xe/ ¹³² Xe	Error	¹²⁶ Xe/ ¹³² Xe	Error	¹²⁸ Xe/ ¹³² Xe	Error	¹²⁹ Xe/ ¹³² Xe	Error
BD11-B	0.003524	027	0.003297	030	0.07073	019	1.0625	026
BD20-B	0.003515	026	0.003277	049	0.07042	016	1.0618	036
WBD04-B	0.003517	028	0.003271	048	0.07054	018	1.0625	018
BD19-B	0.003506	017	0.003249	018	0.07060	011	1.0170	010
BD15-B	0.003517	049	0.003265	053	0.07039	023	1.0333	025
BD02-B	0.003535	019	0.003248	016	0.07054	011	0.9942	014
BD07-B	0.003563	047	0.003310	036	0.07077	025	1.0309	031
BD13-B	0.003553	049	0.003305	046	0.07061	021	1.0135	027
BD05-B	0.003507	024	0.003309	027	0.07090	016	0.9993	021
BD04-B	0.003522	035	0.003279	033	0.07113	017	1.0214	021
WBD01-B	0.003516	022	0.003265	023	0.07061	014	1.0510	018
WBD03-B	0.003542	042	0.003295	039	0.07083	021	1.0490	028
Data from ref. 5	0.003511	021	0.003293	020	0.07044	010	1.0598	005

Sample	¹³⁰ Xe/ ¹³² Xe	Error	¹³¹ Xe/ ¹³² Xe	Error	¹³⁴ Xe/ ¹³² Xe	Error	¹³⁶ Xe/ ¹³² Xe	Error
BD11-B	0.14861	038	0.77556	193	0.41162	103	0.36156	089
BD20-B	0.14783	052	0.77456	291	0.41142	143	0.36195	126
WBD04-B	0.14798	032	0.77568	134	0.41308	078	0.36156	066
BD19-B	0.14927	016	0.78076	076	0.40379	040	0.35017	036
BD15-B	0.14923	041	0.78161	190	0.40905	102	0.35936	092
BD02-B	0.14984	021	0.78354	107	0.40205	055	0.34716	048
BD07-B	0.14943	048	0.78079	233	0.40338	123	0.35296	108
BD13-B	0.14992	042	0.78184	206	0.40535	108	0.35275	095
BD05-B	0.14996	033	0.78240	163	0.40198	085	0.34898	073
BD04-B	0.14919	032	0.77928	161	0.40814	086	0.35580	074
WBD01-B	0.14809	028	0.77369	135	0.41271	073	0.36136	064
WBD03-B	0.14892	045	0.78153	213	0.40997	113	0.36093	099
Data from ref. 5	0.14838	014	0.77549	041	0.41359	027	0.36279	025

Errors are last three digits and 1σ. Included for comparison is previous work⁵.

than the flux out of arcs²⁹. One explanation for this discrepancy is subduction of unbound sea water into the melting zone. If all water fluxing out of arc-related settings is derived from sea water entering the subduction zone together with unfractionated noble gases, this equates to 4.3×10^{28} atoms ³⁶Ar yr⁻¹ available for subduction. Using the ³⁶Ar concentration calculated earlier, and approximating the mass of the convecting mantle to be that of the whole mantle, the convecting mantle contains 6.0×10^{36} atoms ³⁶Ar. With the

assumption that subduction rates have been constant over most of geological time, this means that a subduction efficiency of only 3% over 4.5 Gyr can account for all the ³⁶Ar in the mantle. The convecting mantle value assumes no degassing of the mantle throughout geological time. Taking the extent of mantle degassing to be 40%, as inferred from ⁴⁰Ar in the atmosphere³⁰, increases the subduction efficiency to a maximum of 5%. In either case, if arc-derived water in an unbound form does indeed penetrate into the subduction zone,

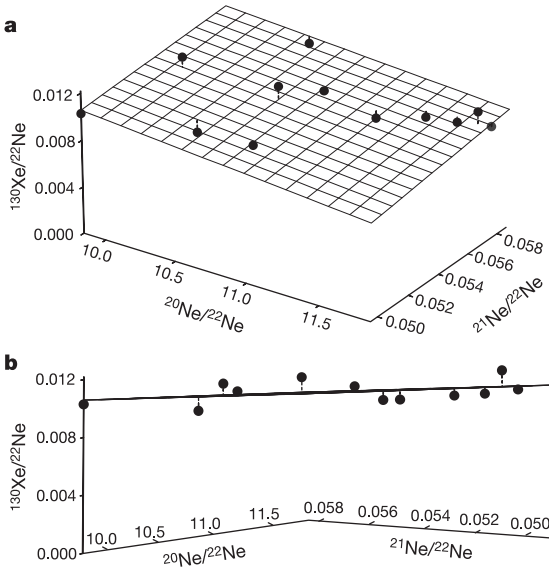


Figure 3 | Correlation of measured ²⁰Ne/²²Ne, ²¹Ne/²²Ne with ¹³⁰Xe/²²Ne. **a**, The plane of best fit to the data. Note the constancy of ¹³⁰Xe/²²Ne in both the groundwater-dominated (²⁰Ne/²²Ne ≈ 9.8) and mantle-rich (²⁰Ne/²²Ne ≈ 11.85) samples. **b**, A rotation of the graph to view the plane edge-on. The good fit to a plane shows that the elemental abundance and isotopic data can be described by the same three end-member components for all samples. Bravo Dome data only are presented.

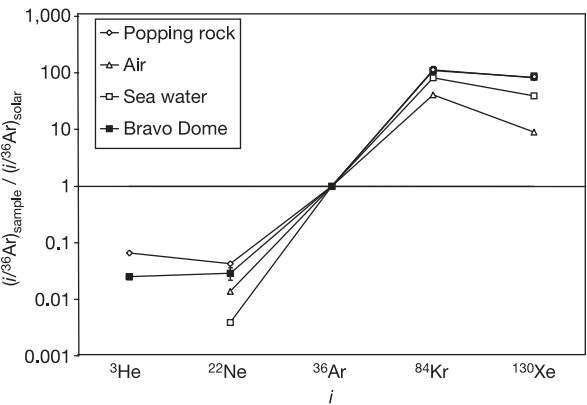


Figure 4 | Mantle noble gas isotopes relative to ³⁶Ar normalized to the solar abundance. The range of values derived for the Bravo Dome system is almost identical to those of the reference mid-ocean-ridge popping rock 2IID43. As both volatile systems are produced by very different fractionating mechanisms, convergence suggests that neither system is significantly fractionated and therefore together provide an unambiguous mantle noble gas abundance pattern. This pattern, when compared with air, suggests that unfractionated air cannot provide the source. The convecting mantle heavy noble gases are clearly most similar to sea water (see text), and this provides *prima-facie* evidence for a seawater origin for these tracers. Error bars include 1σ error of the plane fitted to the data and the uncertainty in the Ne isotopic composition of the mantle end-member (fitting software by J. D. Gilmour, manuscript in preparation).

Table 4 | Resolved mantle elemental ratios

	$^3\text{He}/^{36}\text{Ar}$	$^{22}\text{Ne}/^{36}\text{Ar}$	$^{84}\text{Kr}/^{36}\text{Ar}$	$^{130}\text{Xe}/^{36}\text{Ar}$
Bravo Dome	0.302 (0.024)	0.112 (0.028)	0.0556 (0.0088)	0.001045 (0.00009)
Solar wind	11.92	3.846	0.0005	0.0000125
CI chondrite		0.021	0.0153	0.00700
Air		0.053	0.0207	0.00011
Sea water		0.015	0.0410	0.000563
Popping rock	0.79	0.161	0.0564	0.00103

Popping rock data from ref. 19. Errors in parentheses are 1σ .

accounting for the backarc water flux, <5% of that water is required to progress to the deeper convecting mantle.

Recycling of sea water into the OIB source

There is clear evidence from radiogenic isotope geochemistry that recycled material forms a significant component of OIB material³¹. Similarly, numerical models of whole mantle convection, which do not invoke a physically unobservable chemically derived deep density contrast, show that a significant portion of subducted material reaches the core–mantle boundary, and from this boundary layer contributes to the upwelling plumes³². It is therefore difficult to envisage a mechanism of seawater recycling into the convecting mantle that does not also have a major impact on the OIB source. Although early OIB noble gas isotope data gave results little different to atmospheric Ar and Xe, these have since been shown to be the result of atmospheric contamination³³. The $^{40}\text{Ar}/^{36}\text{Ar}$ ratio of the mantle sourcing plumes is still ambiguous but, nevertheless, the $^{40}\text{Ar}/^{36}\text{Ar}$ upper limit appears to be <10,000, with the Iceland maximum being 4,530 (ref. 34), the Juan Fernandez maximum being 8,000 (ref. 35), the Loihi maximum being 8,300 (ref. 34), and the Reunion maximum being 7,600 (ref. 11). One exception is Samoa, with a $^{40}\text{Ar}/^{36}\text{Ar}$ maximum of 15,000 (ref. 35), but this is probably due to a metasomatic overprint from crustal fluids³⁶. In contrast, the convecting mantle has a far higher proportional contribution from the radiogenic ^{40}Ar isotope, with the $^{40}\text{Ar}/^{36}\text{Ar}$ ratio in the convecting mantle unequivocally resolved to be 35,000–52,500 (ref. 4) and refined with the larger number of mantle-rich samples in this study to $41,050 \pm 2,670$ (Supplementary Information).

A similar difference between OIB and MORB is observed in Xe isotopes. Xe isotope ratios in excesses of air in OIB are only a recent discovery, owing to the high precision required to resolve such a small excess, with $^{129}\text{Xe}/^{130}\text{Xe} < 7$ (refs 34, 37) compared with the air value of 6.49. The convecting mantle value defined by this data set ($^{129}\text{Xe}/^{130}\text{Xe} = 7.90 \pm 0.14$) compares with the highest measured values in MORB popping rock ($^{129}\text{Xe}/^{130}\text{Xe} = 7.5$ (ref. 38) and $^{129}\text{Xe}/^{130}\text{Xe} = 7.73$ (ref. 19)) and from the South Mid-Atlantic Ridge ($^{129}\text{Xe}/^{130}\text{Xe} = 7.78$ (ref. 27)). These observations of low $^{40}\text{Ar}/^{36}\text{Ar}$ and $^{129}\text{Xe}/^{130}\text{Xe}$ isotopic ratios in OIB relative to MORB, combined with the association of OIB with high $^3\text{He}/^4\text{He}$ ratios, have been used to argue that the OIB source is undegassed and therefore a volatile-rich reservoir, preserved since accretion³⁰. We argue that recycling of unfractionated sea water into the OIB source is responsible for the reduction in $^{40}\text{Ar}/^{36}\text{Ar}$ and $^{129}\text{Xe}/^{130}\text{Xe}$ ratios towards air-like values. Deuterium measurements from hotspots are variable, with δD values higher than MORB ranging from -36‰ over the Salas y Gomez plume³⁹ to -50‰ near Iceland⁴⁰. In contrast, lower than MORB values of $\delta\text{D} = -118\text{‰}$ are observed at Loihi, Hawaii⁴¹. In the context of recycling, data with δD lower than MORB reflect a more primitive hydrogen signature present in the lower mantle, and δD higher than MORB is a signature of a proportionally larger recycled component. A more significant recycled seawater contribution to the Iceland source is completely consistent with the lower maximum $^{40}\text{Ar}/^{36}\text{Ar}$ ratios at Iceland ($\sim 4,000$) compared to Hawaii ($\sim 8,000$) and a relatively dry source for the Hawaiian plume⁴².

Recycling dominates the complement of heavy noble gases in the convecting mantle. This clearly precludes ^{36}Ar from being used as a

simple constraint on primitive volatile outgassing from the mantle³⁰. This does not, however, account for the high $^3\text{He}/^4\text{He}$ ratio and higher solar Ne content of many OIBs. It has been shown by several workers that these primitive volatile signatures in OIB can be simply accounted for by the addition of a small volatile-rich component^{43,44}. We speculate that this is associated with OIB during the recycling process by sampling a volatile-rich reservoir preserved at the core–mantle boundary, perhaps at the D'' layer⁴⁵. Alternatively, it has been suggested that the recycling process is not efficient at degassing mantle volatiles, and that the recycled material itself may preserve higher $^3\text{He}/^4\text{He}$ ratios⁴⁶. What is clear is that our explanation for the dichotomy between OIB and MORB heavy noble gas isotope signatures now provides a quantitative constraint on numerical simulations of mantle convection⁴⁷, which in turn may provide a solution to the mechanism of preservation of ancient accretionary volatile signatures in our planet.

Received 11 October 2005; accepted 23 March 2006.

- Lupton, J. E. & Craig, H. Excess ^3He in oceanic basalts, evidence for terrestrial primordial helium. *Earth Planet. Sci. Lett.* **26**, 133–139 (1975).
- Honda, M., Reynolds, J. H., Roedder, E. & Epstein, S. Noble gases in diamonds: occurrences of solar like helium and neon. *J. Geophys. Res.* **92**, 12507–12521 (1987).
- Sarda, P., Staudacher, T. & Allegre, C. J. Neon isotopes in submarine basalts. *Earth Planet. Sci. Lett.* **91**, 73–88 (1988).
- Ballentine, C. J., Sherwood Lollar, B., Marty, B. & Cassidy, M. Neon isotopes constrain convection and volatile origin in the Earth's mantle. *Nature* **433**, 33–38 (2005).
- Caffee, M. W. *et al.* Primordial noble gases from Earth's mantle: identification of a primitive volatile component. *Science* **285**, 2115–2118 (1999).
- Staudacher, T. & Allegre, C. J. Recycling of oceanic crust and sediments: the noble gas subduction barrier. *Earth Planet. Sci. Lett.* **89**, 173–183 (1988).
- Kent, A. J. R. *et al.* Widespread assimilation of a sea-water-derived component at Loihi Seamount, Hawaii. *Geochim. Cosmochim. Acta* **63**, 2749–2761 (1999).
- Kent, A. J. R., Norman, M. D., Hutcheon, I. D. & Stolper, E. M. Assimilation of seawater-derived components in an oceanic volcano: evidence from matrix glasses and glass inclusions from Loihi seamount, Hawaii. *Chem. Geol.* **156**, 299–319 (1999).
- Pepin, R. O. & Porcelli, D. Origin of noble gases in the terrestrial planets. *Rev. Mineral. Geochem.* **47**, 191–246 (2002).
- Kunz, J. Is there solar argon in the Earth's mantle? *Nature* **399**, 649–650 (1999).
- Trieloff, M., Kunz, J. & Allegre, C. J. Noble gas systematics of the Reunion mantle plume source and the origin of primordial noble gases in Earth's mantle. *Earth Planet. Sci. Lett.* **200**, 297–313 (2002).
- Porcelli, D. & Wasserburg, G. J. Mass transfer of xenon through a steady-state upper mantle. *Geochim. Cosmochim. Acta* **59**, 1991–2007 (1995).
- Porcelli, D. & Wasserburg, G. J. Mass transfer of helium, neon, argon, and xenon through a steady-state upper mantle. *Geochim. Cosmochim. Acta* **59**, 4921–4937 (1995).
- Ballentine, C. J. & Barfod, D. N. The origin of air-like noble gases in MORB and OIB. *Earth Planet. Sci. Lett.* **180**, 39–48 (2000).
- Harrison, D., Burnard, P. G., Trieloff, M. & Turner, G. Resolving atmospheric contaminants in mantle noble gas analyses. *Geochim. Geophys. Geosyst.* **4**, art. no. 1023, doi:10.1029/2002GC000325 (2003).
- Sarda, P. & Guillot, B. Breaking of Henry's law for noble gas and CO_2 solubility in silicate melt under pressure. *Nature* **436**, 95–98 (2005).
- Sarda, P. & Graham, D. Mid-ocean ridge popping rocks: implications for degassing at ridge crests. *Earth Planet. Sci. Lett.* **97**, 268–289 (1990).
- Burnard, P., Graham, D. & Turner, G. Vesicle specific noble gas analyses of popping rock: implications for primordial noble gases in Earth. *Science* **276**, 568–571 (1997).
- Moreira, M., Kunz, J. & Allegre, C. J. Rare gas systematics in popping rock: isotopic and elemental compositions in the upper mantle. *Science* **279**, 1178–1181 (1998).

20. Matsuda, J. & Nagao, K. Noble gas abundances in a deep-sea sediment core from eastern equatorial Pacific. *Geochem. J.* **20**, 71–80 (1986).
21. Thompson, A. B. Water in the Earth's upper mantle. *Nature* **358**, 295–302 (1992).
22. Symons, K., Dixon, J., Schilling, J. G., Kingsley, R. & Poreda, R. Volatiles in basaltic glasses from the Easter–Salas y Gomez Seamount Chain and Easter Microplate: implications for geochemical cycling of volatile elements. *Geochem. Geophys. Geosyst.* **3**, doi:10.1029/2001GC000173 (2002).
23. Robert, F., Gautier, D. & Dubrulle, B. The solar system D/H ratio: observations and theories. *Space Sci. Rev.* **92**, 201–224 (2000).
24. Matsumoto, T., Chen, Y. L. & Matsuda, J. Concomitant occurrence of primordial and recycled noble gases in the Earth. *Earth Planet. Sci. Lett.* **188**, 211–219 (2001).
25. Sarda, P., Moreira, M. & Staudacher, T. Argon–lead isotopic correlation in Mid-Atlantic ridge basalts. *Science* **283**, 666–668 (1999).
26. Bach, W. & Niedermann, S. Atmospheric noble gases in volcanic glasses from the southern Lau basin: origin from the subducting slab? *Earth Planet. Sci. Lett.* **160**, 297–309 (1998).
27. Sarda, P., Moreira, M., Staudacher, T., Schilling, J. G. & Allegre, C. J. Rare gas systematics on the southernmost Mid-Atlantic Ridge: constraints on the lower mantle and the Dupal source. *J. Geophys. Res. Solid Earth* **105**, 5973–5996 (2000).
28. Hilton, D. R., Hammerschmidt, K., Looock, G. & Friedrichsen, H. Helium and argon isotope systematics of the central Lau Basin and Valu Fa ridge: evidence of crust–mantle interactions in a back-arc basin. *Geochim. Cosmochim. Acta* **57**, 2819–2841 (1993).
29. Hilton, D. R., Fischer, T. P. & Marty, B. Noble gases and volatile recycling at subduction zones. *Rev. Mineral. Geochem.* **47**, 319–370 (2002).
30. Allegre, C. J., Staudacher, T. & Sarda, P. Rare gas systematics: formation of the atmosphere, evolution and structure of the Earth's mantle. *Earth Planet. Sci. Lett.* **81**, 127–150 (1987).
31. Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
32. Van Keken, P. E., Ballentine, C. J. & Hauri, E. in *Treatise on Geochemistry: The Mantle and Core* (ed. Carlson, R. W.) 471–491 (Elsevier Pergamon, Oxford, 2003).
33. Patterson, D. B., Honda, M. & McDougall, I. Atmospheric contamination—a possible source for heavy noble gases in basalts from Loihi Seamount, Hawaii. *Geophys. Res. Lett.* **17**, 705–708 (1990).
34. Trieloff, M., Kunz, J., Clague, D. A., Harrison, D. & Allegre, C. J. The nature of pristine noble gases in mantle plumes. *Science* **288**, 1036–1038 (2000).
35. Farley, K. A. & Craig, H. Atmospheric argon contamination of ocean island basalt olivine phenocrysts. *Geochim. Cosmochim. Acta* **58**, 2509–2517 (1994).
36. Burnard, P. G., Farley, K. A. & Turner, G. Multiple fluid pulses in a Samoan harzburgite. *Chem. Geol.* **147**, 99–114 (1998).
37. Harrison, D., Burnard, P. & Turner, G. Noble gas behaviour and composition in the mantle: constraints from the Iceland Plume. *Earth Planet. Sci. Lett.* **171**, 199–207 (1999).
38. Kunz, J., Staudacher, T. & Allegre, C. J. Plutonium-fission xenon found in Earth's mantle. *Science* **280**, 877–880 (1998).
39. Kingsley, R. H. et al. D/H ratios in basalt glasses from the Salas y Gomez mantle plume interacting with the East Pacific Rise: water from old D-rich recycled crust or primordial water from the lower mantle? *Geochem. Geophys. Geosyst.* **3**, doi:10.1029/2001GC000199 (2002).
40. Poreda, R., Schilling, J. G. & Craig, H. Helium and hydrogen isotopes in ocean-ridge basalts north and south of Iceland. *Earth Planet. Sci. Lett.* **78**, 1–17 (1986).
41. Hauri, E. SIMS analysis of volatiles in silicate glasses, 2: isotopes and abundances in Hawaiian melt inclusions. *Chem. Geol.* **183**, 115–141 (2002).
42. Dixon, J. E. & Clague, D. A. Volatiles in basaltic glasses from Loihi Seamount, Hawaii: evidence for a relatively dry plume component. *J. Petrol.* **42**, 627–654 (2001).
43. Ballentine, C. J., Van Keken, P. E., Porcelli, D. & Hauri, E. H. Numerical models, geochemistry and the zero-paradox noble-gas mantle. *Phil. Trans. R. Soc. Lond. A* **360**, 2611–2631 (2002).
44. Ellam, R. M. & Stuart, F. M. Coherent He–Nd–Sr isotope trends in high $^3\text{He}/^4\text{He}$ basalts: implications for a common reservoir, mantle heterogeneity and convection. *Earth Planet. Sci. Lett.* **228**, 511–523 (2004).
45. Tolstikhin, I. & Hofmann, A. W. Early crust on top of the Earth's core. *Phys. Earth. Planet. Inter.* **148**, 109–130 (2004).
46. Class, C. & Goldstein, S. L. Evolution of helium isotopes in the Earth's mantle. *Nature* **436**, 1107–1112 (2005).
47. Ballentine, C. J., Van Keken, P. E., Holland, G., Hauri, E. H. & Brandenburg, J. P. Recycling volatiles and attaining a geochemical and fluid dynamically consistent model of mantle convection. *Eos* **86** (52), Fall Meet. Suppl., Abstr. V53F-01 (2005); at (http://www.agu.org/meetings/fm05/fm05-sessions/fm05_V53F.html) (2005).
48. York, D. Least squares fitting of a straight line with correlated errors. *Earth Planet. Sci. Lett.* **5**, 320–324 (1969).
49. Ott, U. Noble gases in meteorites—trapped components. *Rev. Mineral. Geochem.* **47**, 71–100 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Oxy for permission to sample the Bravo Dome Field, Amerada Hess for permission to sample West Bravo Dome, and M. Cassidy for logistics and field assistance. We thank D. Blagburn and B. Clementson for laboratory support. We also thank J. Gilmour for provision of the error weighted plane fitting software, S. Gilfillan for air calibration work and D. Hilton for comments on the manuscript. This work was funded by NERC.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.H. (g.holland@manchester.ac.uk).

LETTERS

Neptune's capture of its moon Triton in a binary-planet gravitational encounter

Craig B. Agnor¹ & Douglas P. Hamilton²

Triton is Neptune's principal satellite and is by far the largest retrograde satellite in the Solar System (its mass is ~ 40 per cent greater than that of Pluto). Its inclined and circular orbit lies between a group of small inner prograde satellites and a number of exterior irregular satellites with both prograde and retrograde orbits. This unusual configuration has led to the belief that Triton originally orbited the Sun before being captured in orbit around Neptune^{1–3}. Existing models^{4–6} for its capture, however, all have significant bottlenecks that make their effectiveness doubtful. Here we report that a three-body gravitational encounter between a binary system (of $\sim 10^3$ -kilometre-sized bodies) and Neptune is a far more likely explanation for Triton's capture. Our model predicts that Triton was once a member of a binary with a range of plausible characteristics, including ones similar to the Pluto–Charon pair.

One possible outcome of gravitational encounters between a binary system and a planet is an exchange reaction, where one member of the binary is expelled and its place taken by the planet (Fig. 1). Analogous three-body encounters have been studied in a variety of contexts^{7–10}; these studies suggest that the process may be relevant to the capture of planetary satellites¹⁰ in general and for Triton¹¹ in particular. Here we develop an analytic description of this process and evaluate it using numerical simulations of binaries encountering Neptune. Satellite capture by this pathway requires that: (1) the capture candidate be a member of a binary, (2) the binary be disrupted during the encounter, and (3) one of its members be left permanently bound to the planet.

Binaries have recently been discovered in nearly all of the Solar System's small-body reservoirs and appear to be a natural consequence of planet formation and Solar System evolution^{9,12–16}. Recent surveys have found satellites orbiting $\sim 16\%$ of near-Earth asteroids¹⁷, $\sim 2\%$ of large main-belt asteroids¹⁸, and $\sim 11\%$ of Kuiper-belt objects¹⁹, including Pluto and the recently discovered 2003 UB₃₁₃. Given the observational constraints, the true fraction of objects with satellites is probably larger, and binary–planet encounters are therefore highly probable.

Three-body encounters will render the binary unbound when its centre of mass passes close enough to the planet that the binary separation is approximately its Hill radius. This occurs at a tidal disruption distance of

$$r_{\text{td}} = a_B \left(\frac{3M_p}{m_1 + m_2} \right)^{1/3} = R_p \left(\frac{a_B}{R_1} \right) \left[\left(\frac{3\rho_p}{\rho_1} \right)^{1/3} \left(\frac{m_1}{m_1 + m_2} \right)^{1/3} \right] \quad (1)$$

from the planet, where a_B is the binary semi-major axis, M_p , R_p and ρ_p are the planet mass, radius and density, respectively, and $m_{1,2}$ and ρ_1 are the binary masses and density with $m_2 < m_1$. As the term in square brackets is near unity, the disruption distance measured in planetary radii is nearly the binary's separation measured in radii of its largest component (R_1). Numerical simulations of binary–planet encounters confirm this as the effective scaling length for binary

disruption. Results also show disruption to be a strong function of the inclination of the binary orbit relative to the encounter plane (I_B). Prograde binaries with $I_B < 90^\circ$ are efficiently disrupted for close approach distances $q_e \lesssim r_{\text{td}}$, but retrograde ones ($I_B > 90^\circ$) require much deeper encounters. Similar inclination dependence of disruption has been observed in several studies of tidal interactions^{10,20,21}.

Using the results of simulated binary–planet encounters as a guide, we find that a simple model, which assumes that the binary is impulsively disrupted, provides an effective description of the gravitationally focused encounters with Neptune considered here (Figs 2 and 3). As the binary approaches the planet on a hyperbolic trajectory, m_1 and m_2 orbit their mutual centre of mass (Fig. 1). On disruption, the smaller body (m_2) experiences a change in speed of the order of its orbital speed about the binary centre of mass

$$\Delta v_2 \approx \pm \frac{m_1}{(m_1 + m_2)} \left(\frac{G(m_1 + m_2)}{a_B} \right)^{1/2} \quad (2)$$

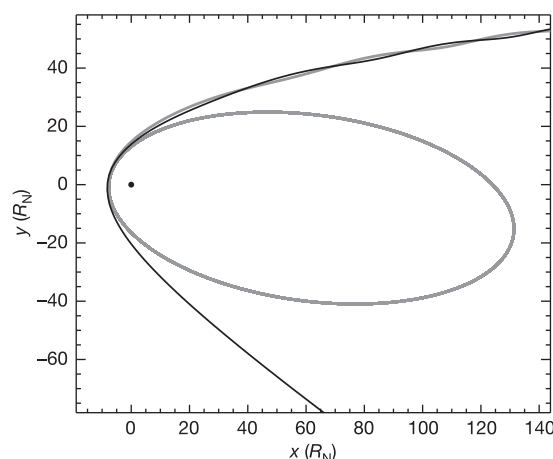


Figure 1 | Exchange capture of Triton. We show trajectories of an example encounter in arbitrary planetocentric cartesian coordinates between an equal-mass Triton binary ($m_1 = m_2 = m_T$, where m_T is Triton's mass) and Neptune. The encounter is planar ($I_B = 0$) with encounter speed at infinity $v_\infty = 0.50 \text{ km s}^{-1}$ and close approach distance $q_e = 8R_N$, where R_N is Neptune's radius. The initial binary orbit is circular ($e_B = 0.0$) with semi-major axis $a_B = 20R_1 \approx 1.1R_N$, where R_1 is the radius of the larger component, and we have assumed a density of $\rho_1 = 2.0 \text{ g cm}^{-3}$ (similar to Triton). The binary approaches from the upper right and is disrupted while interior to $\sim r_{\text{td}} = 21R_N$ from Neptune. One member is captured to an orbit with semi-major axis $a_c \approx 70R_N$ while the other escapes. We used a Bulirsch-Stoer integrator to model binary–planet gravitational encounters. For a given set of encounter dynamics (v_∞ , q_e) and binary characteristics ($m_{1,2}$, a_B , e_B , I_B), we performed hundreds of simulations with different initial binary orbital phases.

¹Earth Sciences Department, Center for the Origin, Dynamics and Evolution of Planets, 1156 High Street, University of California, Santa Cruz, California 95064, USA. ²Department of Astronomy, University of Maryland at College Park, College Park, Maryland 20742-2421, USA.

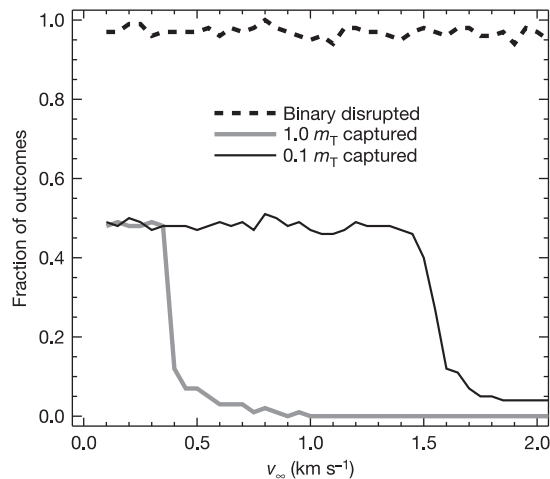


Figure 2 | Outcomes of simulated binary-planet encounters. Simulation results of a Triton binary ($m_1 = m_T$, $m_2 = 0.1m_T$, $a_B = 20R_N$, $e_B = 0.0$) encountering Neptune at speeds $0 < v_\infty < 2.0 \text{ km s}^{-1}$ are shown. Larger values are unlikely, as Neptune's orbital speed is only $v_N = 5.4 \text{ km s}^{-1}$, and crossing orbits have relative speeds $v_\infty \approx e_\odot v_N$, where $e_\odot$ is the heliocentric eccentricity. In all cases, $I_B = 0^\circ$ and the binary centre-of-mass close approach distance is $q_c = 8R_N$. The encounters are deep within the binary tidal radius ($r_{td} = 26R_N$), and permanent disruption occurs in $\sim 95\%$ of cases (the dashed black line). As the binary tumbles towards the planet, each component spends half its orbit moving faster and half slower than the binary centre of mass. The more massive object (m_1) moves with a smaller, but non-negligible velocity relative to the centre of mass (see equation (2)). For $v_\infty \leq 0.35 \text{ km s}^{-1}$, either binary member is captured with roughly 50% probability owing to orbital phasing at the time of disruption. At higher velocity ($0.35 < v_\infty \leq 1.55 \text{ km s}^{-1}$) capture of m_1 is possible, but rare, while m_2 is still captured at a 50% rate owing to its greater orbital speed in the binary centre-of-mass frame.

where G is the gravitational constant and $\Delta v_1 = \Delta v_2 m_2 / m_1$. When combined with energy arguments, this change in speed allows the new semi-major axes to be determined (Fig. 3). Because the tidal forces that cause binary disruption are maximized when the three bodies are most nearly collinear, preferred binary orbital phases (and values of Δv_i near that of equation (2)) are selected when disruption occurs. This results in a clustering of capture semi-major axes (a_c) around a characteristic value (Fig. 3).

When the binary orbital angular momentum is much less than the encounter angular momentum, as in all cases considered here, the pericentre of the capture orbit (q_c) is comparable to that of the planet–centre of mass close approach distance (q_c) offset by the binary orbital separation, (that is, $q_c \approx q_e - a_B m_1 / (m_1 + m_2)$ for capture of m_2 with $I_B < 90^\circ$; Fig. 1). As the inclination of the capture orbit is nearly that of the centre-of-mass trajectory about the planet, capture to prograde or retrograde orbits about Neptune is determined by this path. In principle, this mechanism can transfer an object to virtually any satellite orbit if the requirements of disruption and capture can be satisfied by the encounter dynamics (q_e , v_∞ , I_B) and the binary characteristics ($m_{1,2}$, a_B). In practice, highly elliptical capture orbits are favoured. We are examining the role of three-body encounters in the capture and evolution of additional relevant populations; a report is forthcoming.

The capture of Triton must transfer it to an orbit contained within Neptune's Hill sphere ($r_H = a_N (M_N / 3M_\odot)^{1/3}$ where M_N , a_N are Neptune's mass and semi-major axis respectively, and $M_\odot$ is the solar mass) and the semi-major axis of the capture orbit is limited to values $a_c < r_H / 2 = 2,300R_N$. This large eccentric orbit may have intersected and/or strongly perturbed regions where Triton's neighbouring satellites now reside, driving catastrophic collisions between primordial regular satellites and/or exciting the orbits of irregular satellites via close encounters and secular interactions^{1,6,22,23}. Following

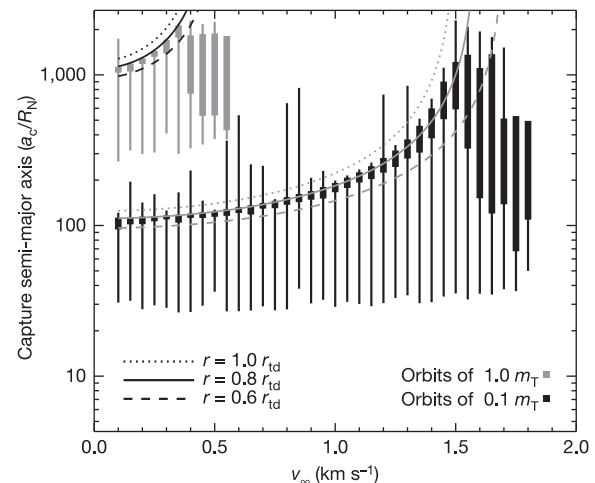


Figure 3 | Determining capture orbits. For the simulations in Fig. 2, we show the range of orbital semi-major axes of the capture orbits (a_c normalized to Neptune's radius, R_N) as a function of the encounter speed (v_∞). Thin vertical lines connect the minimum and maximum values of the a_c for the larger ($m_1 = m_T$; grey) and smaller ($m_2 = 0.1m_T$; black) binary components observed in our simulations. We have omitted these lines for encounters where less than 5% of simulations resulted in capture. Thicker lines denote the range where 50% of the capture orbits lie, and indicate that the distribution of semi-major axes is clustered toward a characteristic value for each encounter velocity. When $M_p \gg m_{1,2}$, impulsive transfer from a hyperbolic encounter orbit with approach speed v_∞ to a bound elliptical orbit with semi-major axis a_c at a distance r from the planet requires a reduction in speed of $\Delta v_c = \sqrt{v_\infty^2 + 2GM_p/r} - \sqrt{GM_p/(2r - 1/a_c)}$. Equating $\Delta v_c = \Delta v_2$ from equation (2) yields an expression relating the binary characteristics to the encounter and capture orbits. Overplotted curves show the values of a_c predicted assuming that a characteristic reduction in speed from equation (2) occurs at distances of $r = 0.6, 0.8$ and $1.0r_{td}$ from the planet, with $r = 0.8r_{td}$ providing a good fit to the median a_c . For other binary characteristics and encounter dynamics, we usually find good fits in the range $r = 0.06 - 0.8r_{td}$. Results for modestly inclined ($I_B \lesssim 60^\circ$) and/or eccentric binaries are qualitatively similar.

capture, Triton's orbital semi-major axis and eccentricity decayed to the currently observed $a_T = 14.06R_N$ and $e_T = 4 \times 10^{-4}$, owing to satellite tides^{1,24} and accumulation of collisional debris²². Tides alone cause the orbit to damp while maintaining constant angular momentum, suggesting that the pericentre of Triton's eccentric capture orbit was near $q_c \approx a_T/2$. Capture to this q_c requires close passage to Neptune with $q_e \lesssim 7R_N$ and binary separations and masses such that $q_e \lesssim r_{td}$. We note that for large capture orbits (for example, $a_c \gtrsim 200R_N$) solar and secular perturbations can drive substantial oscillations in Triton's orbital angular momentum^{6,22,23,25}, expanding the range of plausible capture pericentres considerably.

Using these constraints, we show that Triton's capture can be facilitated by binaries with a wide range of characteristics (Fig. 4). The semi-major axes must satisfy $(a_B/R_N) \gtrsim 5$ for efficient disruption, so that $r_{td} > 7R_N$, and Triton's companion must be sufficiently massive to provide the required impulse. In practice this is not particularly restrictive, and the escaping object can actually be less massive than Triton. For small values of v_∞ and large capture orbits a_c , the escaping companion may be as small as a few times $0.01m_T$.

Previous models for Triton's capture invoke aerodynamic drag in a protosatellite gas disk⁵ or collision with a pre-existing regular satellite of Neptune⁶. For aerodynamic drag, capture must be carefully timed, occurring during the lifetime of the protosatellite gas disk—of the order of $10^3 - 10^6 \text{ yr}$ (ref. 3)—and just before the disk's dispersal to avoid continued orbital decay into the planet. Collision capture requires an extremely unlikely event. The probability of colliding with a single regular satellite given one encounter with $q_e < 10R_N$ is

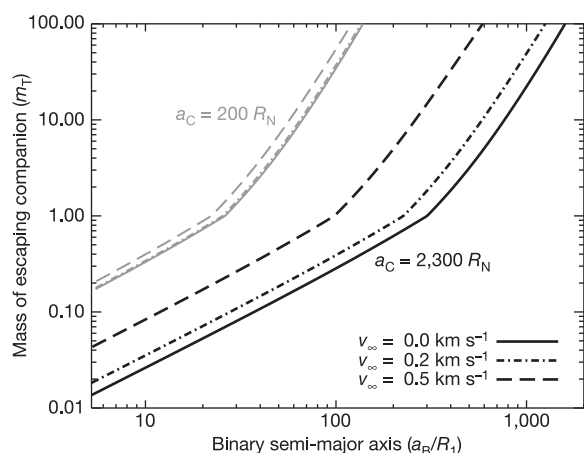


Figure 4 | Binaries capable of delivering Triton to Neptune. Efficient binary disruption requires $q_c < r_{td}$; using equations (1) and (2), and assuming that the change in speed occurs near $r = 0.8r_{td}$, we solve for the mass of the escaping companion. We show the masses (in units of Triton's mass, m_T) required for capture with $a_c \leq 2,300 R_N$ (black) and with $a_c = 200 R_N$ (grey) as a function of a_B/R_1 (where a_B is the semi-major axis of the binary, and R_1 is the radius of the largest component of the binary). These curves were generated by inverting fits to the median values in Fig. 3, and are representative of a range of the binary parameters needed to accomplish the transition of Triton between encounters with speed v_∞ and capture orbits of size a_c . Scenarios of Neptune's accumulation^{26,29,30} give encounter speeds $v_\infty < 0.5 \text{ km s}^{-1}$ or $e_0 < 0.1$. Delivery to small semi-major axes is possible (for example, Fig. 1) and becomes more probable with more massive companions. Binaries facilitating Triton's capture have properties similar to those of known Kuiper belt binaries (that is, $m_2/m_1 \approx 0.05$ – 1 , $a_B/R_1 \approx 17$ – $1,300$)¹⁸ and to the Pluto–Charon pair ($m_{\text{Charon}}/m_{\text{Pluto}} = 1/8$ and $a_B/R_{\text{Pluto}} \approx 17$) in particular.

low ($P_c \approx (R_T)^2/(10R_N)^2 \approx 3 \times 10^{-5}$, where R_T is Triton's radius), requiring numerous ($\approx 10^3$) close passes with Neptune to make the occurrence of a single collision reasonable. Also, the striking satellite must be large enough to capture, but not so large ($\approx 0.01m_T$) as to catastrophically disrupt, Triton^{3,26}. A review of these models³ slightly favours collisional capture. Clearly, both mechanisms require specialized conditions, ones primarily available during Neptune's formation, to function.

Gravitational capture of one binary component offers a number of significant advantages. As with previous models, capture can be realized for conditions prevalent during Neptune's formation. However, capture might also occur later as encounters between Neptune and material from an exterior planetesimal disk drove the planet's outward migration^{27,28}. Further, exchange capture is gentle and brief for Triton and does not concomitantly subject it to loss via collisional disruption or continued orbital decay. Depending on the binary characteristics, Triton may be captured to a wide range of satellite orbits, including those tightly bound to Neptune ($a_c \lesssim 100R_N$; see Fig. 1 for an example), making binary exchange capture consonant with many different plausible orbital histories for Triton and evolutionary paths of the Neptune satellite system.

Consider a single binary of the right type (Fig. 4) approaching Neptune with $q_e < r_{td}$. Assuming an isotropic distribution of binary orientations (I_B), 25% will have encounters with $I_B \leq 60^\circ$, and will be efficiently disrupted and captured. Noting the additional 50% chance of securing Triton from the binary, the odds that this single binary–Neptune encounter will result in Triton's capture are at least 1/8.

Exchange capture is favoured over collision capture if the probability that Triton once resided in a favourable binary exceeds $\sim 8P_c \approx 3 \times 10^{-4}$ (assuming similar capture cross-sections, that is $q_e \leq 10R_N$). Given the observed 11% lower limit on the frequency of Kuiper-belt object binaries¹⁹ and the weak constraints on the binary characteristics required, we find exchange capture much more likely

than collision⁶. We conclude that Triton was once a member of a binary and was captured as it made a close approach to Neptune.

Received 29 December 2005; accepted 31 March 2006.

- McCord, T. B. Dynamical evolution of the Neptunian system. *Astron. J.* **71**, 585–590 (1966).
- McKinnon, W. B. On the origin of Triton and Pluto. *Nature* **311**, 355–358 (1984).
- McKinnon, W. B., Lunine, J. I. & Banfield, D. *Neptune and Triton* (ed. Cruikshank, D. P.) 807–877 (Univ. Arizona Press, Tucson, 1995).
- Pollack, J. B., Burns, J. A. & Tauber, M. E. Gas drag in primordial circumplanetary envelopes—A mechanism for satellite capture. *Icarus* **37**, 587–611 (1979).
- McKinnon, W. B. & Leith, A. C. Gas drag and the orbital evolution of a captured Triton. *Icarus* **118**, 392–413 (1995).
- Goldreich, P., Murray, N., Longaretti, P. Y. & Banfield, D. Neptune's story. *Science* **245**, 500–504 (1989).
- Heggie, D. C. Binary evolution in stellar dynamics. *Mon. Not. R. Astron. Soc.* **173**, 729–787 (1975).
- Hills, J. G. Computer simulations of encounters between massive black holes and binaries. *Astron. J.* **102**, 704–715 (1991).
- Funato, Y., Makino, J., Hut, P., Kokubo, E. & Kinoshita, D. The formation of Kuiper-belt binaries through exchange reactions. *Nature* **427**, 518–520 (2004).
- Morbidelli, A. & Levison, H. F. Scenarios for the origin of the orbits of the trans-neptunian objects 2000 CR₁₀₅ and 2003 VB₁₂ (Sedna). *Astron. J.* **128**, 2564–2576 (2004).
- Tsui, K. H. Satellite capture in a four-body system. *Planet. Space Sci.* **50**, 269–276 (2002).
- Durda, D. D. et al. The formation of asteroid satellites in large impacts: results from numerical simulations. *Icarus* **170**, 243–257 (2004).
- Weidenschilling, S. J. On the origin of binary transneptunian objects. *Icarus* **160**, 212–215 (2002).
- Goldreich, P., Lithwick, Y. & Sari, R. Formation of Kuiper-belt binaries by dynamical friction and three-body encounters. *Nature* **420**, 643–646 (2002).
- Astakhov, S. A., Lee, E. A. & Farrelly, D. Formation of Kuiper-belt binaries through multiple chaotic scattering encounters with low-mass intruders. *Mon. Not. R. Astron. Soc.* **360**, 401–415 (2005).
- Canup, R. M. A giant impact origin of Pluto–Charon. *Science* **307**, 546–550 (2005).
- Margot, J. L. et al. Binary asteroids in the near-earth object population. *Science* **296**, 1445–1448 (2002).
- Merline, W. J. et al. *Asteroids III* (eds Bottke, W. F. Jr, Cellino, A., Paolicchi, P. & Binzel, R. P.) 289–312 (Univ. Arizona Press, Tucson, 2002).
- Stephens, D. C. & Noll, K. S. Detection of six trans-neptunian binaries with NICMOS: A high fraction of binaries in the cold classical disk. *Astron. J.* **131**, 1142–1148 (2006).
- Hamilton, D. P. & Burns, J. A. Orbital stability zones about asteroids. II—The destabilizing effects of eccentric orbits and of solar radiation. *Icarus* **96**, 43–64 (1992).
- Asphaug, E. & Benz, W. Size density, and structure of comet Shoemaker–Levy 9 inferred from the physics of tidal breakup. *Icarus* **121**, 225–248 (1996).
- Čuk, M. & Gladman, B. J. Constraints on the orbital evolution of Triton. *Astrophys. J.* **626**, L113–L116 (2005).
- Hamilton, D. P., Zhang, K. & Agnor, C. B. Constraints on Triton's orbital evolution. *AAS/Div. Dyn. Astron. Meet.* **36**(2), 11.04 (2005).
- Goldreich, P. & Soter, S. Q in the Solar System. *Icarus* **5**, 375–389 (1966).
- Benner, L. A. M. & McKinnon, W. B. Orbital behaviour of captured satellites: The effect of solar gravity on Triton's postcapture orbit. *Icarus* **114**, 1–20 (1995).
- Lissauer, J. J., Pollack, J. B., Wetherill, G. W. & Stevenson, D. J. *Neptune and Triton* (ed. Cruikshank, D. P.) 37–108 (Univ. Arizona Press, Tucson, 1995).
- Hahn, J. M. & Malhotra, R. Orbital evolution of planets embedded in a planetesimal disk. *Astron. J.* **117**, 3041–3053 (1999).
- Gomes, R. S., Morbidelli, A. & Levison, H. F. Planetary migration in a planetesimal disk: why did Neptune stop at 30 AU? *Icarus* **170**, 492–507 (2004).
- Ward, W. R., Agnor, C. B. & Tanaka, H. in *Astrophysical Ages and Time Scales* (eds von Hippel, T., Simpson, C. & Manset, N.) 111–120 (ASP Conf. Ser. 245, Astronomical Society of the Pacific, San Francisco, 2001).
- Goldreich, P., Lithwick, Y. & Sari, R. Planet formation by coagulation: a focus on Uranus and Neptune. *Annu. Rev. Astron. Astrophys.* **42**, 549–601 (2004).

Acknowledgements We thank E. Asphaug for advice and support and C. McGleam for a critical reading of the manuscript. This work was supported by the National Aeronautics and Space Administration under grants issued through the Planetary Geology and Geophysics, Outer Planets Research, and Origins of Solar Systems programmes.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.B.A. (cagnor@pmc.ucsc.edu).

Local switching of two-dimensional superconductivity using the ferroelectric field effect

K. S. Takahashi¹, M. Gabay², D. Jaccard¹, K. Shibuya³, T. Ohnishi³, M. Lippmaa³ & J.-M. Triscone¹

Correlated oxides display a variety of extraordinary physical properties including high-temperature superconductivity¹ and colossal magnetoresistance². In these materials, strong electronic correlations often lead to competing ground states that are sensitive to many parameters—in particular the doping level—so that complex phase diagrams are observed. A flexible way to explore the role of doping is to tune the electron or hole concentration with electric fields, as is done in standard semiconductor field effect transistors³. Here we demonstrate a model oxide system based on high-quality heterostructures in which the ferroelectric field effect approach can be studied. We use a single-crystal film of the perovskite superconductor Nb-doped SrTiO₃ as the superconducting channel and ferroelectric Pb(Zr,Ti)O₃ as the gate oxide. Atomic force microscopy is used to locally reverse the ferroelectric polarization, thus inducing large resistivity and carrier modulations, resulting in a clear shift in the superconducting critical temperature. Field-induced switching from the normal state to the (zero resistance) superconducting state was achieved at a well-defined temperature. This unique system could lead to a field of research in which devices are realized by locally defining in the same material superconducting and normal regions with 'perfect' interfaces, the interface being purely electronic. Using this approach, one could potentially design one-dimensional superconducting wires, superconducting rings and junctions, superconducting quantum interference devices (SQUIDs) or arrays of pinning centres.

Undoped perovskite SrTiO₃ is a band insulator. On chemical doping with oxygen, Nb or La (SrTiO_{3-x}, SrTi_{1-x}Nb_xO₃, Sr_{1-x}La_xTiO₃), the system undergoes a series of transitions from an insulating to a semiconducting state and, finally, to a metallic state. At low temperatures and at very low carrier densities between about 1×10^{19} and 10^{21} cm^{-3} , the system becomes superconducting⁴⁻⁶. Here, atomically flat Nb-doped SrTiO₃ films are used as an oxide superconductor. As SrTiO₃ is the 'natural' substrate for many perovskites, including ferroelectric Pb(Zr,Ti)O₃, single-crystal Pb(Zr,Ti)O₃ films can be grown on the Nb-doped SrTiO₃ films with a high degree of perfection. This allows us to explore the ferroelectric field effect approach^{3,7,8} in this ideal low carrier density system. Epitaxial heterostructures composed of a 500-Å-thick layer of ferroelectric Pb(Zr_{0.2}Ti_{0.8})O₃ (PZT) and a 260-Å-thick layer of superconducting Sr(Ti_{0.98}Nb_{0.02})O₃ (Nb-STO) were fabricated on (001) SrTiO₃ (STO) single-crystal substrates. As schematically illustrated in Fig. 1a, standard photolithography and wet etching processes were used to pattern the devices in a geometry suitable for transport measurements. Experiments below 4.2 K and down to 30 mK were performed using a dilution refrigerator. The metallic tip of an atomic force microscope (AFM), acting as a mobile gate electrode, was used to sweep the entire area of the conducting path (the blue box region shown in Fig. 1a) at room temperature, causing a local switching of

the direction of the ferroelectric polarization—parallel or anti-parallel to the *c* axis. The topographic image measured at the same time is shown in Fig. 1b. After the poling scan, AFM piezoresponse was used to determine the ferroelectric domain structure. A phase image, corresponding to a map of the material piezoelectric response is shown in Fig. 1c, revealing the two polarization states (P+ (dark red) and P− (light red)) and the artificially modified domain structure.

Figure 2a shows the temperature dependence of the resistivity for the two polarization states. The P+ state, poled using −12 V (on the tip), corresponds to the polarization direction that removes electrons from the Nb-STO layer. The P− state, poled using +12 V, corresponds to the polarization state that adds electrons to the Nb-STO layer, effectively increasing the doping level. Across the entire range of

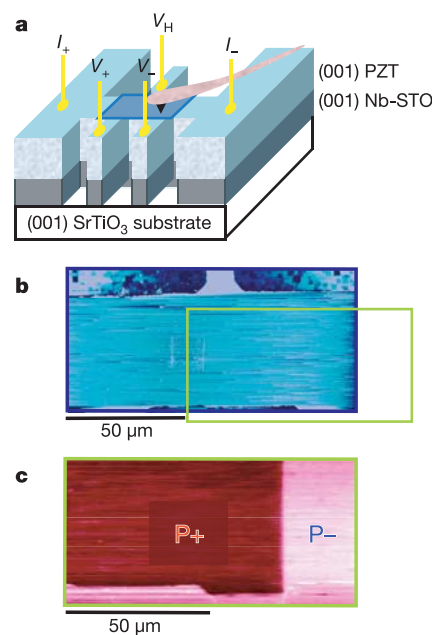


Figure 1 | Diagram of the device, and topographic and piezoelectric images of the field effect active region. **a**, Diagram of the device. The ferroelectric polarization was switched in the blue box area using the AFM tip as a local electrode, scanning the whole area. PZT, Pb(Zr_{0.2}Ti_{0.8})O₃; Nb-STO, Sr(Ti_{0.98}Nb_{0.02})O₃; V_H, V₊ and V_− are the voltage electrodes for Hall and resistivity measurements; I₊ and I_− are the current electrodes. **b**, AFM topographic image measured during the poling scan (120 μm × 60 μm). **c**, Piezoresponse image of the green rectangle area in **b** after poling from the P− state to the P+ state. (P+ corresponds to the polarization direction that removes electrons from the Nb-STO layer. In the P− state, the polarization field adds electrons to the film.)

¹DPMC, University of Geneva, 24 Quai Ernest Ansermet, 1211 Geneva 4, Switzerland. ²Laboratoire de Physique des Solides, Université de Paris-Sud, 91405 Orsay, France.

³Institute for Solid State Physics, University of Tokyo, Chiba 277-8581, Japan.

our temperature scan, both states showed metallic behaviour and displayed large (field induced) resistance modulations. As expected from the electron-like conduction character of doped STO, the resistivity of the P− state is lower than that of the P+ state, as the field effect modulates the electron carrier density. At room temperature, a 60% difference between the two resistivity states was observed, and this value gradually decreased down to 20% as the temperature was reduced.

In order to investigate quantitatively the ferroelectric field effect carrier modulation, the Hall effect was measured for the two polarization states at various temperatures. Figure 2b shows the inverse Hall coefficient (R_H^{-1}) for the two states and the difference (ΔR_H^{-1}) as a function of temperature. As the field effect induces a carrier profile at the interface in the Nb-STO layer, the measured value of R_H^{-1} is somehow 'averaged' over the whole Nb-STO layer thickness (260 Å). The characteristic width of the carrier modulation length in a semiclassical metallic system is given by the Thomas–Fermi screening length: $\lambda_{TF} = (\epsilon E_F / 6\pi n e^2)^{1/2}$, where ϵ is the dielectric constant, E_F the Fermi energy, n the carrier density and e the electronic charge. For doped STO ($n = 5 \times 10^{19} \text{ cm}^{-3}$, $E_F = 30 \text{ meV}$), λ_{TF} at 300 K is $\sim 30 \text{ Å}$ and λ_{TF} below 5 K is $\sim 210 \text{ Å}$, assuming that the temperature dependence of ϵ is similar to that of undoped single crystals⁹. From 260 K to 100 K, a very large, almost temperature independent, modulation of the Hall response ($\sim 60\%$) was observed, its size being comparable to that of the corresponding modulation in resistivity. Below 100 K, however, R_H^{-1} dramatically decreases for both polarization states. This behaviour is very different from the essentially temperature independent Hall coefficient observed in single crystals¹⁰, making the large temperature dependence observed for ΔR_H^{-1} below 100 K very intriguing.

In a simple single band model, ΔR_H^{-1} would be proportional to the difference in carrier density (Δn), $\Delta n = \Delta P / ed$, where ΔP is the change in the PZT polarization and d is the thickness of the Nb-STO layer (260 Å). Because ΔP is essentially constant below 300 K (the ferroelectric Curie temperature of PZT being $\sim 700 \text{ K}$), one would expect ΔR_H^{-1} to be constant¹¹. Given that the Hall coefficient is

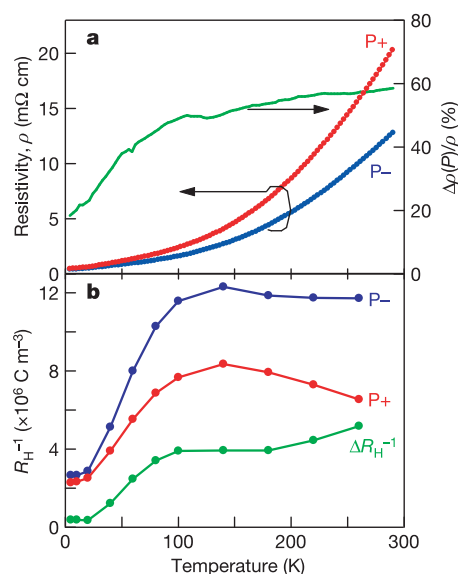


Figure 2 | Transport properties for the two polarization states. **a**, Left axis: temperature dependence of the resistivity (ρ) for the two polarization states, P+ and P−. Right axis (green curve): temperature dependence of the resistivity difference ratio between the two polarization states defined as $\Delta\rho(P)/\rho = (\rho(P+) - \rho(P-))/\rho(P-)$. **b**, Temperature dependence of the inverse Hall coefficient (R_H^{-1}) and the difference $\Delta R_H^{-1} = R_H^{-1}(P-) - R_H^{-1}(P+)$.

temperature independent above 100 K and below 20 K, and that T_c (the superconducting critical temperature) versus doping thin film data (discussed later) can only be compared with single crystals using the high temperature Hall effect data, one could infer that the strong decrease in inverse Hall coefficient does not reflect carrier freezing but may rather be related to localization, band effects or multiband conduction. To determine the carrier densities n (calculated from $R_H^{-1} = en$), we have thus used the high temperature, T independent, inverse Hall coefficient. This gives us an average value of about $(4-5) \times 10^{19} \text{ cm}^{-3}$ for the P+ state and of $7.5 \times 10^{19} \text{ cm}^{-3}$ for the P− state. These values correspond to approximately 20% of active Nb sites, a fraction reasonably close to reports on thin films and single crystals^{8,12,13}. Although the exact doping mechanism is unknown, Nb doping turned out to be very robust. Annealing at 500 °C only marginally changes the film conductivity, and experiments carried out over two years did not affect the properties of the devices. The measured change in carrier density implies a ΔP value of $10-14.5 \mu\text{C cm}^{-2}$ ($P \approx 6 \mu\text{C cm}^{-2}$). This value is of the order of the polarization field measured in similar devices, suggesting that the high temperature Hall coefficient yields a correct estimate of the carrier density.

Figure 3a shows the temperature dependence of the resistivity of each state below 380 mK. Both states display sharp superconducting transitions, reaching zero resistance, with a clear shift of the transition observed. With T_c defined as the temperature at which the

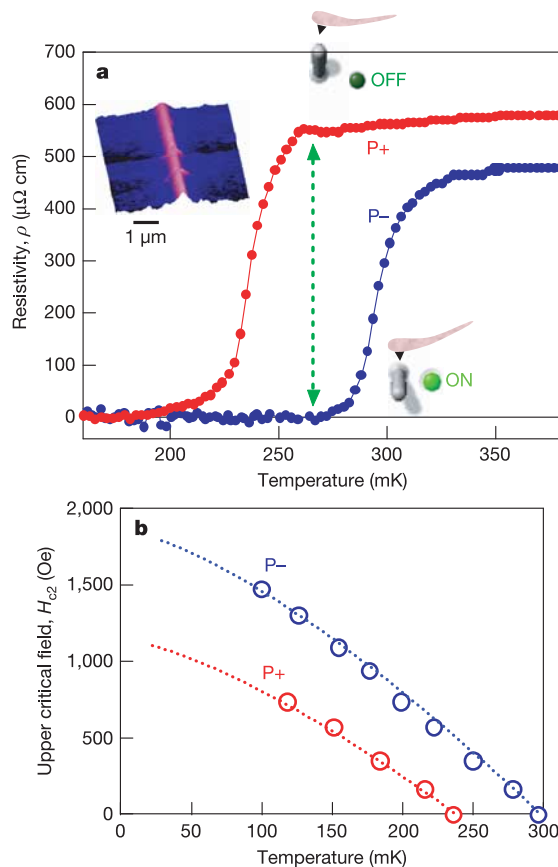


Figure 3 | Superconducting properties for the two polarization states.

a, Resistivity versus temperature at low temperatures. As indicated by the arrow and the associated insets, a superconducting on-off switching is observed around 270 mK, illustrated by the 'AFM switch' shown as insets. The left inset shows the piezoresponse image after poling a P+ state line in the P− state background. **b**, Temperature dependence of the upper critical field, H_{c2} , for each state. $T_c(H)$ is defined as the temperature at which the resistivity decreases to 50% of the resistivity at 400 mK. Dotted lines indicate the WHH model fittings (see text).

resistivity decreases to 50% of the value at 400 mK, the T_c value for the P− state is 296 mK and that of the P+ state is 236 mK. Around 270 mK, the green arrow shows that switching of the ferroelectric polarization induces a transition from the normal state (P+ state) to a zero resistance superconducting state (P− state). In general, such a ‘complete’ switching is very difficult to obtain, because transitions to the superconducting state are often broad in very thin films (in fact broader than the induced shift of T_c). In high- T_c oxide films^{14,15} and in ultrathin bismuth films¹⁶, a field induced superconductor–insulator transition was however observed in few cases. The complete switching observed here is extremely promising for the promotion of the local ferroelectric field effect idea proposed in 1997 by Ahn *et al.*⁷, and it could lead to novel nanoscale superconducting devices. As an example, the inset of Fig. 3a shows the piezoresponse image of part of the conducting path after AFM writing of a P+ domain line (red) in the P− background (blue).

Superconductivity is suppressed by applying a magnetic field along the [001] direction (out of the film plane). Figure 3b shows the upper critical field H_{c2} as a function of T . $T_c(H)$ is defined as the temperature at which the resistivity drops down to 50% of its value at 400 mK (we will justify this choice below). The temperature dependence of $H_{c2}(T)$ can be roughly fitted to the WHH model that describes conventional dirty limit type-II superconductors¹⁷, in agreement with an estimation of ξ_0/l (~ 30) $\gg 1$, where ξ_0 is the BCS superconducting coherence length¹⁹ and l the normal state mean free path. The estimated upper critical field at $T=0$ is about 2,000 Oe for the P− state and 1,100 Oe for the P+ state, respectively. The superconducting coherence length can thus be calculated using $H_{c2}(0) = \Phi_0/2\pi\xi(0)^2$, where Φ_0 is the flux quantum and $\xi(0)$ is the coherence length at 0 K, which gives 400 Å for the P− state and 550 Å for the P+ state. If the scattering rate is taken to be the same for the P+ and P− states, the change in coherence length can be related to the changes in T_c , resistivity and Hall coefficient, leading to a 33% change as compared to the 40% experimentally observed. The fact that the coherence length is larger (for both polarization states) than the film thickness (260 Å) indicates that the superconducting system is two dimensional (2D).

The 2D nature of the superconducting transition can be confirmed by an analysis of paraconductivity fluctuations. Resistivity measurements with low noise level reveal that the resistivity starts to deviate slightly from the normal state T behaviour around 1.4 K with a temperature dependence that can be fitted to a 2D fluctuation model^{18,19}. A fit of resistivity, ρ , versus T for the P+ and P− states allows us to determine the mean field critical temperature of each state. These values match closely the values of T_c obtained from the ‘50% of the normal state resistance’ criterion, thus ‘justifying’ a posteriori our T_c determination and validating our H_{c2} analyses. Of course T_c plays the role of a mean field critical temperature as the actual zero field transition is expected to be Kosterlitz–Thouless (KT) like²⁰. Fitting the tails of the resistive transitions yields estimates of the transition temperature T_{KT} and of the mean field T_c (ref. 21). For the P+ and the P− states we get T_{KT} values of 170 mK and 252 mK, and T_c values of 239 mK and 295 mK, respectively. Using Beasley *et al.*’s criterion²² also allows T_{KT} and T_c to be determined. Making use of the relationship $T_{KT}/T_c = (1 + 0.173R_{\square}/(\hbar/e^2))^{-1}$ where $R_{\square} = \rho_N/d$ is the sheet resistance, ρ_N the normal state resistance and d the film thickness, one would conclude that $T_{KT} \approx T_c$ for the two states, which clearly disagrees with our findings.

Gabay *et al.* claimed²³ that the smallness of the core energy invalidates the standard KT scenario for most superconducting films, as it leads to large effective vortex dielectric constants, ϵ^* . Instead, they predicted the existence of a new vortex-antivortex lattice phase, which melts via a KT transition at a temperature T_{KT} sizably smaller than that derived by Beasley *et al.* Because the transition temperature is much less than T_c , the superfluid sheet density can be expected to approach its asymptotic zero temperature limit and T_{KT} should be proportional to the carrier sheet density so

long as ϵ^* is not too strongly T dependent. This is illustrated in Fig. 4, which shows the dependence of the mean field (Fig. 4a) and KT transition temperatures (Fig. 4b) (determined above and listed in the figure legend) on carrier density n for the two states mapped onto the ‘chemical doping’ STO phase diagram. We identify the T_{KT} that we extracted from our resistivity curves with the melting temperature of the vortex-antivortex crystal. As can be seen in Fig. 4a and b, the T_c values for the two states fall reasonably well on the reduced STO curve, while T_{KT} is proportional to n as illustrated by the black dotted line. As the coherence length of superconducting Nb-STO is larger than d , although the change in carrier density is induced only near the Nb-STO/PZT interface, the Cooper pairs experience an average effective interaction through the proximity effect. We thus plot the average carrier densities calculated from the average $R_{\square}^{-1}$ in Fig. 4. Note that using the low temperature Hall coefficient values would lead to points that lie far off the single crystal curves. It is noteworthy that the field induced modulations of T_{KT} and T_c reflect different physics. T_{KT} , as explained above, is directly linked to the 2D superfluid density which is modified by the field effect. On the other hand, the electrostatic modulation of the mean field T_c is consistent with

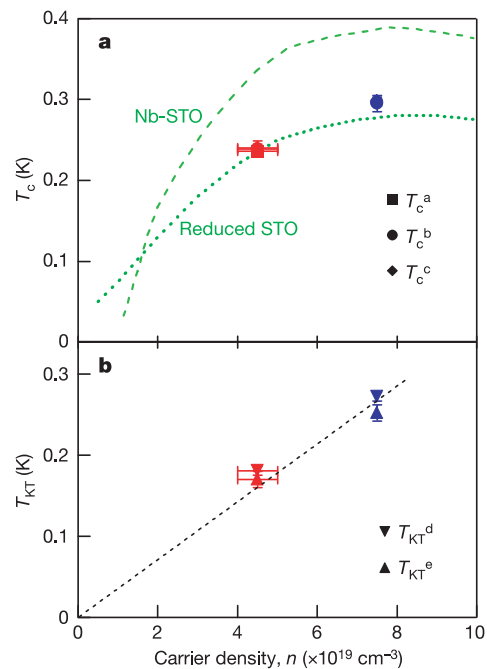


Figure 4 | Dependence of the critical temperature on carrier density for the two polarization states. **a**, The dependence of the mean field transition temperatures, T_c , on carrier density, n , for the two polarization states. The behaviour of reduced STO bulk crystals (green dotted line) and Nb-doped STO bulk crystals (green dashed line) are also included. **b**, The dependence of the Kosterlitz–Thouless transition temperatures (T_{KT} ; see text) on n for the two polarization states; the dotted line is a guide to the eye to highlight the linear relationship. Values of n are obtained from the high temperature Hall coefficient values. The T_c criteria (T_c^a and T_c^d) and the values for the two polarization states are as follows. T_c^a , the temperature at which the resistivity decreases to 50% of that at 400 mK, that is, Beasley’s criterion (P+ and P−, 236 mK and 296 mK); T_c^b , the mean field transition temperature determined by fitting the bottom part of the resistivity curves²¹ (239 mK, 295 mK); T_c^c , the mean field transition temperature determined by fitting the onset of the transitions with a 2D fluctuation model (240 mK, 298 mK); T_{KT}^d , the temperature at which the resistivity decreases to 1% of that at 400 mK, that is, Beasley’s criterion for T_{KT} (181 mK, 272 mK); T_{KT}^e , the KT transition temperature determined by fitting the bottom part of resistivity curves²¹ (170 mK, 252 mK). The error bars on T_c and T_{KT} result from uncertainties on the fitting. The maximum error is estimated to be about 10 mK. The error bars on n correspond to the observed small variations in the Hall coefficient at high temperatures.

the well known (but yet to be understood) bell-shape T_c phase diagram illustrated by the dotted and dashed lines in Fig. 4a.

In conclusion, we have achieved large ferroelectric field effects using a model system device based on thin films of PZT and Nb-STO. Large modulations of the resistivity, inverse Hall coefficient and 2D superconductivity are obtained. At a given temperature, complete switching from the normal state to a zero resistance superconducting state upon reversal of the PZT polarization direction is observed, an effect that represents a new step towards nanoscale electronic devices.

METHODS

Heterostructures. Epitaxial heterostructures composed of 500-Å-thick PZT and 260-Å-thick Nb-STO were fabricated on TiO₂-terminated (001) STO single-crystal substrates.

Nb-STO layer deposition. A pulsed laser deposition method employing KrF excimer laser pulses (100 mJ) focused on a commercial 1 wt% (2 atm%) Nb-doped SrTiO₃ single-crystal target was used. During deposition, the substrate temperature was kept between 1,200 and 1,300 °C and the oxygen pressure between 2×10^{-7} and 2×10^{-6} torr. In order to get conductive films, high substrate temperature and low oxygen pressure during the deposition turned out to be necessary. An infrared (807 nm) diode laser was used to reach the required high substrate temperature, illuminating the substrate from outside the ultra-high-vacuum chamber through a view port²⁴. By monitoring the intensity of reflection high-energy electron diffraction (RHEED) during the deposition, step-flow type growth was observed. No particles were observed even in large areas (typically $5 \times 5 \mu\text{m}^2$), a clear advantage for the subsequent growth of high quality epitaxial PZT. After deposition, the films were cooled in 760 torr of oxygen to avoid reducing the film and the substrate.

PZT layer deposition. This layer was deposited by off-axis radio frequency magnetron sputtering in an argon–oxygen flow (Ar:O₂ = 3:1) at a total pressure of 0.18 torr and at a substrate temperature of ~500 °C. From X-ray diffraction analyses of these bilayers, the PZT layers are found to be *c*-axis oriented and single-crystal like.

Ferroelectric polarization. To switch the direction of the ferroelectric polarization (parallel or anti-parallel to the *c* axis), a constant voltage of +12 V or –12 V, leading to a field that is larger than the coercive field of PZT, was applied between the tip and the Nb-STO layer. To determine the ferroelectric domain structure, the AFM piezoresponse imaging mode was used. In this mode, the PZT layer is subjected to an a.c. (10 kHz) applied field (1.5 V) and the local linear piezoelectric response is measured with lock-in detection. The sign of the piezoelectric deformation reveals the sign of the ferroelectric polarization. In Fig. 1c, the dark and light red correspond to each sign of the ferroelectric polarization. During the transport measurements, there was no application of any gate voltage, making the device insensitive to possible leakage currents.

Received 11 October 2005; accepted 16 March 2006.

- Bednorz, J. G. & Müller, K. A. Possible high- T_c superconductivity in the Ba-La-Cu-O system. *Z. Phys. B* **64**, 189–193 (1986).
- Tokura, Y. (ed.) *Advances in Condensed Matter Science* Vol. 2, *Colossal Magnetoresistive Oxides* (Gordon and Breach, London, 2000).
- Ahn, C. H., Triscone, J.-M. & Mannhart, J. Electric field effect in correlated oxide systems. *Nature* **424**, 1015–1018 (2003).
- Schooley, J. F., Hosler, W. R. & Cohen, M. L. Superconductivity in semiconducting SrTiO₃. *Phys. Rev. Lett.* **12**, 474–475 (1964).
- Pfeiffer, E. R. & Schooley, J. F. Superconducting transition temperatures of Nb-doped SrTiO₃. *Phys. Lett. A* **29**, 589–590 (1969).
- Koonce, C. S., Cohen, M. L., Schooley, J. F., Hosler, W. R. & Pfeiffer, E. R. Superconducting transition temperatures of semiconducting SrTiO₃. *Phys. Rev.* **163**, 380–390 (1967).
- Ahn, C. H. *et al.* Local, nonvolatile electronic writing of epitaxial Pb(Zr_{0.52}Ti_{0.48})O₃/SrRuO₃ heterostructures. *Science* **276**, 1100–1103 (1997).
- Takahashi, K. S. *et al.* Electrostatic modulation of the electronic properties of Nb-doped SrTiO₃ superconducting films. *Appl. Phys. Lett.* **84**, 1722–1724 (2004).
- Müller, K. A. & Burkard, H. SrTiO₃: An intrinsic quantum para-electric below 4 K. *Phys. Rev. B* **19**, 3593–3602 (1979).
- Tufte, O. N. & Chapman, P. W. Electron mobility in semiconducting strontium titanate. *Phys. Rev.* **155**, 796–802 (1967).
- Aidam, R., Fuchs, D. & Schneider, R. Ferroelectric field effect in YBa₂Cu₃O_{7- δ} thin films. *Physica C* **328**, 21–30 (1999).
- Lee, C., Destry, J. & Brebner, J. L. Optical absorption and transport in semiconducting SrTiO₃. *Phys. Rev. B* **11**, 2299–2309 (1975).
- Leitner, A. *et al.* Pulsed laser deposition of superconducting Nb-doped strontium titanate thin films. *Appl. Phys. Lett.* **72**, 3065–3067 (1998).
- Ahn, C. H. *et al.* Electrostatic modulation of superconductivity in ultrathin GdBa₂Cu₃O_{7- δ} films. *Science* **284**, 1152–1155 (1999).
- Cassinese, A., Luca, G. M. D., Prigibbo, A., Salluzzo, M. & Vaglio, R. Field-effect tuning of carrier density in Nd_{1.2}Ba_{1.8}Cu₃O_y thin films. *Appl. Phys. Lett.* **84**, 3933–3935 (2004).
- Parendo, K. A. *et al.* Electrostatic tuning of the superconductor-insulator transition in two dimensions. *Phys. Rev. Lett.* **94**, 197004 (2005).
- Werthamer, N. R., Helfand, E. & Hohenberg, P. C. Temperature and purity dependence of superconducting critical field H_{c2} . III. Electron spin and spin-orbit effects. *Phys. Rev.* **147**, 295–302 (1966).
- Takahashi, K. S. *et al.* Epitaxial growth and transport properties of Nb-doped SrTiO₃ thin films. *Proc. SPIE* **5932**, 267–274 (2005).
- Tinkham, M. *Introduction to Superconductivity* 2nd edn (McGraw-Hill, New York, 1996).
- Kosterlitz, J. M. & Thouless, D. J. Long range order and metastability in two dimensional solids and superfluids. *J. Phys. C* **5**, L124–L126 (1972).
- Minnhagen, P. The two-dimensional Coulomb gas, vortex unbinding, and superfluid-superconducting films. *Rev. Mod. Phys.* **59**, 1001–1066 (1987).
- Beasley, M. R., Mooji, J. E. & Orlando, T. P. Possibility of vortex-antivortex pair dissociation in two-dimensional superconductors. *Phys. Rev. Lett.* **42**, 1165–1168 (1979).
- Gabay, M. & Kapitulnik, A. Vortex-antivortex crystallization in thin superconducting and superfluid films. *Phys. Rev. Lett.* **71**, 2138–2141 (1993).
- Ohashi, S. *et al.* Compact laser molecular beam epitaxy system using laser heating of substrate for oxide film growth. *Rev. Sci. Instrum.* **70**, 178–183 (1999).

Acknowledgements We thank M. Dawber for a careful reading of the manuscript. This work was supported by the Swiss National Science Foundation through the National Center of Competence in Research, 'Materials with Novel Electronic Properties, MaNEP' and division II, New Energy and Industrial Technology Development Organization (NEDO) of Japan, and ESF (Thiox).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-M.T. (Jean-Marc.Triscone@physics.unige.ch).

Strained silicon as a new electro-optic material

Rune S. Jacobsen¹, Karin N. Andersen¹†, Peter I. Borel¹, Jacob Fage-Pedersen¹, Lars H. Frandsen¹, Ole Hansen², Martin Kristensen¹†, Andrei V. Lavrinenko¹, Gaid Moulin¹, Haiyan Ou¹, Christophe Peucheret¹, Beáta Zsigri¹ & Anders Bjarklev¹

For decades, silicon has been the material of choice for mass fabrication of electronics. This is in contrast to photonics, where passive optical components in silicon have only recently been realized^{1,2}. The slow progress within silicon optoelectronics, where electronic and optical functionalities can be integrated into monolithic components based on the versatile silicon platform, is due to the limited active optical properties of silicon³. Recently, however, a continuous-wave Raman silicon laser was demonstrated⁴; if an effective modulator could also be realized in silicon, data processing and transmission could potentially be performed by all-silicon electronic and optical components. Here we have discovered that a significant linear electro-optic effect is induced in silicon by breaking the crystal symmetry. The symmetry is broken by depositing a straining layer on top of a silicon waveguide, and the induced nonlinear coefficient, $\chi^{(2)} \approx 15 \text{ pm V}^{-1}$, makes it possible to realize a silicon electro-optic modulator. The strain-induced linear electro-optic effect may be used to remove a bottleneck⁵ in modern computers by replacing the electronic bus with a much faster optical alternative.

The inversion symmetry of a non-strained silicon crystal prohibits the existence of a linear electro-optic effect³. However, we have found that the symmetry can be broken by applying an asymmetric strain to the crystal by depositing a straining layer on top of it (Fig. 1), hence creating a linear electro-optic effect in silicon. In other words, when silicon is properly strained, its bulk refractive index (n) varies linearly as a function of external applied electric field (E). In the present structure, we used silicon nitride glass (Si_3N_4) as a straining layer⁶. The amorphous Si_3N_4 layer is compressively strained, that is, it tries to expand the structure underneath in both horizontal directions.

The ability to induce a change in the refractive index can be applied to make an amplitude modulator using a Mach–Zehnder interferometer (Fig. 2), where an applied electric field determines whether or not incident light is transmitted through the modulator. Such an

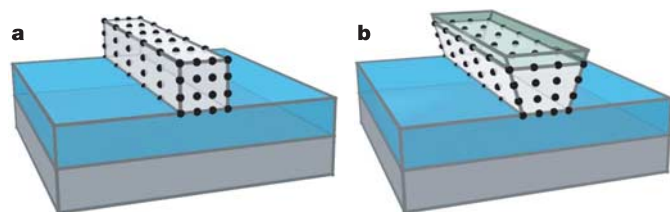


Figure 1 | Applying strain to crystalline silicon. **a**, Waveguide fabricated in the top layer of an SOI wafer. **b**, The same waveguide with a straining layer deposited on top. The straining layer breaks the inversion symmetry and induces a linear electro-optic effect.

electro-optic modulator⁷ is typically used when transmitting data, where transmitted light corresponds to a ‘1’ bit and no light to a ‘0’ bit.

The effect of the material nonlinearity⁸ $\chi^{(2)}$ is enhanced if the waveguide possesses a permanently enlarged group index, n_g . It is theoretically predicted⁹ that the enhancement scales linearly with the group index. Here, the enhanced nonlinear coefficient $\chi_{\text{enh}}^{(2)}$ is therefore defined as:

$$\chi_{\text{enh}}^{(2)} \equiv \frac{n_g}{n} \chi^{(2)} \quad (1)$$

An enlarged n_g can for instance be obtained by guiding the light in a photonic crystal waveguide¹⁰ (PCW), which is a waveguide surrounded by a periodically microstructured material^{11–13} (Fig. 2). This is the approach that was followed in this work when the strain-induced nonlinearity was discovered. Here, the microstructure consists of silicon with holes that are partially filled with glass and air, giving rise to large material index variations. The period of the structure is on the same scale as the wavelength of the incident light. The result is a waveguide with extremely high group indices—for example, reaching values above 230 for the investigated structure¹⁴.

The group index of a PCW is highly polarization-dependent, and in the present work only transverse electric (TE) light is considered. The experimental method applied to ensure TE polarization of the light transmitted through the PCW is described in detail elsewhere¹⁴. As the electric modulation field is applied in the x -direction (the

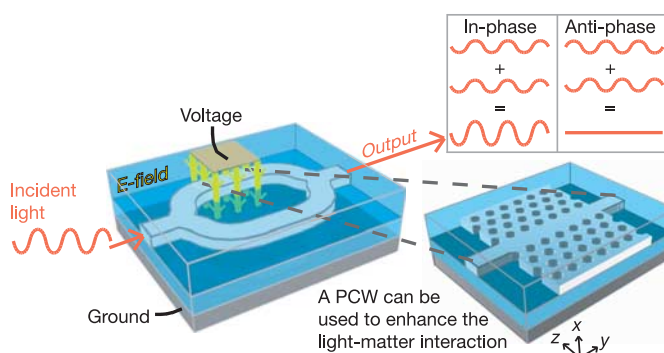


Figure 2 | Diagram of a Mach–Zehnder modulator. Incident light is split into two waveguides. The output amplitude depends on the phase difference at recombination. As shown at top right, in-phase recombination gives a ‘1’ bit output while anti-phase recombination gives a ‘0’ bit output. If the straight waveguide in the Mach–Zehnder modulator is replaced with a PCW as illustrated with the magnified waveguide at bottom right, the effect of a material nonlinearity can be enhanced.

¹COM-DTU, Department of Communications, Optics & Materials, Building 345V, Nano-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. ²CINF, MIC-Department of Micro and Nanotechnology, Building 345E, Nano-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. †Present addresses: MIC, Building 345E, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark (K.A.); iNANO and IFA, University of Aarhus, Ny Munkegade, Building 1520, DK-8000 Århus C, Denmark (M.K.).

coordinate system is shown in Fig. 2) and the light is polarized in the z -direction (TE-polarized), the measured element of the $\chi^{(2)}$ tensor is $\chi_{zzx}^{(2)}$. In the investigated structure, a spacing layer of silica glass (SiO_2) was sandwiched between the silicon waveguide and the straining Si_3N_4 glass (Fig. 3a). By this arrangement, the guided light is confined well below the Si_3N_4 straining layer; that is, the optical field does not extend to this layer. This arrangement ensures that the only influence of the Si_3N_4 layer on the silicon waveguide is a physical strain, exerted through the SiO_2 layer. By measuring $\chi_{\text{enh},zzx}^{(2)}$ both before and after deposition of the Si_3N_4 straining layer, it is demonstrated that the observed linear electro-optic effect is strain-induced.

The measurement before deposition of the straining Si_3N_4 was performed at wavelengths with particularly high n_g values of ~ 70 and ~ 180 to ensure optimal detection. A maximum $\chi_{\text{enh},zzx}^{(2)}$ of $\sim 60 \text{ pm V}^{-1}$ was found (Fig. 3b) for the highest n_g value at a wavelength of $1,561.468 \pm 0.005 \text{ nm}$. The existence of this small but non-zero $\chi_{\text{enh},zzx}^{(2)}$ before deposition of the straining layer is due to a small tensile strain in the upper SiO_2 layer—the SiO_2 spacing layer tries to contract in the horizontal directions. After deposition of the straining Si_3N_4 layer, a $\chi_{\text{enh},zzx}^{(2)}$ of $\sim 830 \text{ pm V}^{-1}$ was found at exactly the same wavelength ($1,561.468 \pm 0.005 \text{ nm}$), so the straining layer increased $\chi_{\text{enh},zzx}^{(2)}$ in the silicon by one order of magnitude (Fig. 3b).

The difference in the induced nonlinear coefficient before and after depositing the Si_3N_4 layer is not only caused by the different magnitudes of strain, but also by the different signs of the strain. The upper SiO_2 layer tries to contract, which is only possible if the whole structure bends like a deep plate. This deformation is only minor, as the thick supporting wafer is extremely difficult to bend. After depositing the Si_3N_4 layer, the sign of the strain is changed, and the top layers now try to expand in the horizontal directions. This expansion can be partially accommodated by a slight wafer bending and partially by introducing strong local deformations near

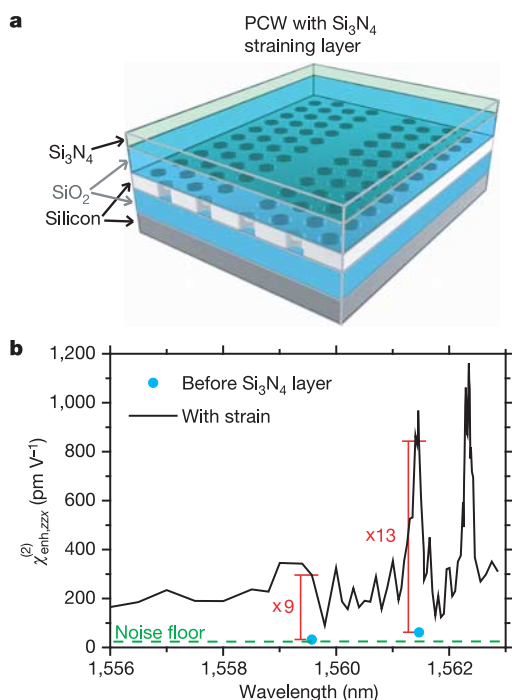


Figure 3 | Effect of straining the silicon structure. **a**, Diagram of the investigated PCW. **b**, Enhanced nonlinearity $\chi_{\text{enh},zzx}^{(2)}$ measured before (blue points) and after (black line) applying the straining layer. A strain-induced increase in $\chi_{\text{enh},zzx}^{(2)}$ by one order of magnitude is observed, compared to the weakly strained structure before deposition of Si_3N_4 . The noise floor is shown as the dashed green line.

each hole. As a result, the compressive strain is more effective in deforming the waveguiding silicon layer. A Si_3N_4 top layer is of course not required, as strain can also be obtained by altering the deposition conditions of the SiO_2 layer. In fact, another structure was fabricated, in which the SiO_2 layer deposited directly on top of the PCW was grown with a surplus of silicon radicals. Without any additional layers, a $\chi_{\text{enh},zzx}^{(2)}$ nonlinearity was induced. The size of this nonlinearity was about two-thirds of what was obtained when the Si_3N_4 straining layer was used.

By comparing $\chi_{\text{enh},zzx}^{(2)}$ with an independent measurement of the group index (Fig. 4), it is confirmed that the wavelength-dependent changes of $\chi_{\text{enh},zzx}^{(2)}$ are caused by dispersion of the group index. This is, to our knowledge, the first clear experimental verification of the theoretically predicted⁹ linear relationship (equation (1)) between group index and enhanced nonlinear coefficient. A previous experiment¹⁵ has shown the existence of an enhancement effect, but it did not verify the predicted linear relationship. To consolidate this result, it can be added that the measured group index was found to be in excellent agreement with both two- and three-dimensional finite-difference time domain (FDTD) calculations performed using the time-of-flight principle. Both the group index measurement and the FDTD calculations are described in detail elsewhere¹⁴.

Several experiments were performed to verify that the observed change in refractive index is due to a genuine $\chi_{zzx}^{(2)}$ rather than being caused by other previously reported effects in silicon¹⁶ (that is, by thermo-optic, photo-elastic, photo-electric or interface effects, or by an effective $\chi_{zzx}^{(2)}$ originating from $\chi_{zzx}^{(3)}$).

First, to rule out a thermo-optic and/or a photo-elastic effect, it is noted that both effects are independent of the sign of the electric field. Hence, for these effects the response for an applied sinusoidal modulation, $\sin(\omega t)$, is $|\sin(\omega t)|$, which has a large frequency component at 2ω . However, no refractive index change was observed at the frequency 2ω . Second, to show that an effective second-order nonlinearity $\chi_{\text{eff},zzx}^{(2)}$, originating from the combination of $\chi_{zzx}^{(3)}$ and a frozen-in electric field¹⁷, does not contribute, $\chi_{\text{enh},zzx}^{(2)}$ was measured while applying a large d.c. field ($> 40 \text{ V } \mu\text{m}^{-1}$ over the silicon layer) across the structure. The value of $\chi_{\text{enh},zzx}^{(2)}$ was unaffected by the size and sign of the applied d.c. field, thus excluding a $\chi_{\text{eff},zzx}^{(2)}$ contribution. We note that in the investigated PCW structure, carriers can move in the horizontal direction in the waveguiding layer and thereby move in and out of the silicon region underneath the electrode. The experiment that excluded a $\chi_{\text{eff},zzx}^{(2)}$ contribution also demonstrates that a possible field-induced change in carrier concentration does not play a role, as the large applied positive d.c.

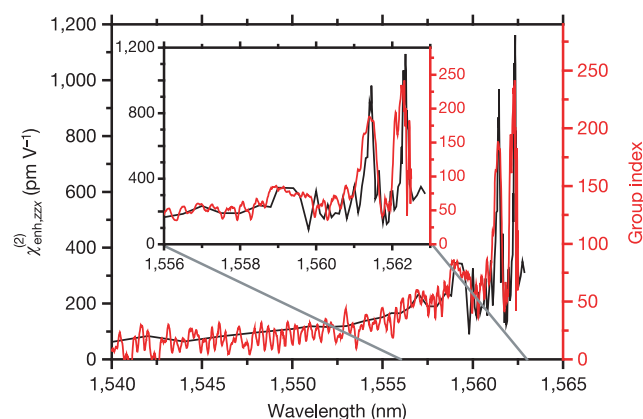


Figure 4 | Experimental verification of the linear enhancement of $\chi_{\text{enh},zzx}^{(2)}$ with the group index. The enhanced nonlinear coefficient $\chi_{\text{enh},zzx}^{(2)}$ (black) of the strained silicon is plotted together with the group index (red) of the photonic crystal waveguide. The inset shows the details in the wavelength range where the fluctuations of the group index are largest.

field efficiently depleted the silicon for carriers and, hence, made carrier transportation impossible.

Third, to investigate if carriers generated by the photo-electric effect contributed to an index change, the experiment with an applied d.c. field was repeated in the dark, ensuring that no photo-generated carriers were present. The photon energy of the incident laser light at $\sim 1,550$ nm is lower than the silicon bandgap, that is, the possible carrier contribution from this light is negligible as it only arises from higher-order effects, like two-photon absorption. The result of the measurement was unaffected by the applied d.c. field even when no carriers were generated by light absorption. Hence, the observed refractive index change does not originate from a photo-electric effect.

Fourth, to evaluate the existence of possible effects originating from the interfaces between different materials, a new structure was fabricated. A ~ 1 - μm -wide single-mode rib waveguide was defined in a similar silicon-on-insulator (SOI) wafer with a ~ 340 -nm-thick top silicon layer, by etching only ~ 30 nm down on both sides of the waveguide core. The fabrication methods applied for deposition of the spacing SiO_2 and the straining Si_3N_4 layers on the PCW were reused to deposit identical layers on top of this shallow etched waveguide. The waveguide fabricated in this way has the same interfaces as the described PCW structure, but the lack of holes in the waveguiding silicon layer hinders local spatial deformation of the silicon crystal. To obtain a large light-matter interaction for such a waveguide, possessing a normal group index ($n_g \approx 3.6$), a 500 times longer (1 cm) waveguide was characterized. No $\chi_{zzx}^{(2)}$ was observed, that is, the $\chi_{zzx}^{(2)}$ value for this structure was below 0.02 pm V^{-1} . This finally substantiates that the observed change in refractive index is caused by a genuine $\chi_{zzx}^{(2)}$ effect originating from deformation of the silicon crystal.

In order to compare the strain-deformed silicon material with other known nonlinear materials, one must distinguish between the enhanced nonlinearity $\chi_{\text{enh}}^{(2)}$ and the material nonlinearity $\chi^{(2)}$. The magnitude of $\chi_{\text{enh}}^{(2)}$ is governed both by the material's nonlinear coefficient and by the group index (equation (1)), where the group index is determined by the design of the waveguide. To make a fair comparison between different nonlinear materials, one must compare quantities that are independent of the waveguide design, that is, one must compare the material $\chi^{(2)}$ values. The material $\chi_{zzx}^{(2)}$ value is determined to be $\sim 15 \text{ pm V}^{-1}$ for the examined strain-deformed silicon from the relationship $\chi^{(2)} = (n/n_g)\chi_{\text{enh}}^{(2)}$, by inserting the bulk refractive index of silicon ($n = 3.5$) and the two independent measurement curves for $\chi_{\text{enh},zzx}^{(2)}$ and for n_g (Fig. 4).

It is relevant to compare the size of the strain-induced material nonlinearity in silicon ($\sim 15 \text{ pm V}^{-1}$) with that of the commonly applied nonlinear material, LiNbO_3 . In LiNbO_3 , the largest $\chi^{(2)}$ tensor component⁷ ($\chi_{zzz}^{(2)}$) is $\sim 360 \text{ pm V}^{-1}$. However, the required electrode spacing must be considered, as a design aim for a modulator is to reduce the required driving voltage, not the electric field. The large vertical index step in an SOI waveguide (for example, in a photonic wire¹⁸) allows an ~ 9 times smaller electrode spacing compared to a LiNbO_3 waveguide¹⁹, and hence the difference in required driving voltage is reduced accordingly. Moreover, the material $\chi^{(2)}$ value of strained silicon reported here may increase significantly if the silicon crystal is allowed to deform more freely, using a photonic wire (Fig. 1) instead of a PCW. One of the current advantages of LiNbO_3 is the low loss of 0.2 dB cm^{-1} obtainable in a waveguide²⁰. Currently the lowest reported optical propagation loss in a silicon photonic wire²¹ is 0.8 dB cm^{-1} . However, the loss in a photonic wire is caused by surface roughness²¹ and it can be reduced by further process development. The advantage of silicon in comparison with LiNbO_3 is the possibility of applying cheap standard fabrication techniques used for fabrication of CMOS circuits. The silicon modulator can also be integrated monolithically with an electronic driver circuit and hybridized with existing semiconductor lasers.

The strain-induced nonlinearity is not the only possibility for realizing an optical modulator in silicon. A modulator based on changing the carrier concentration has been demonstrated²², and a significant increase of the modulation speed up to 10 Gbit s^{-1} has recently been obtained^{23,24}. However, changing the carrier concentration requires a large a.c. current²⁴ of $\sim 0.2 \text{ A}$ (r.m.s.) through the silicon structure. In contrast, in the strained silicon presented here, there is no current through the waveguide structure. Moreover, the speed of the strained silicon modulator is not limited by charge mobility or charge recombination times. Thus, the discovered linear electro-optic effect may provide a decisive step towards utilization of active silicon-based photonics.

METHODS

Sample fabrication. A 20- μm -long straight W1 PCW was fabricated²⁵ using CMOS technology to etch holes with a diameter of $255 \pm 5 \text{ nm}$ placed in a triangular pattern with pitch $\Lambda = 435 \text{ nm}$ in the top layer of a SOI wafer. The waveguide was defined in the ΓK direction by removing one row of holes. The SOI wafer consisted of $220 \pm 20 \text{ nm}$ single crystalline silicon on $1,000 \pm 22.5 \text{ nm}$ SiO_2 , placed on a single crystalline silicon wafer. The silicon was doped with boron, and the resistivity of the top and bottom $\langle 100 \rangle$ oriented silicon layers are $8.5\text{--}11.5$ and $14\text{--}18.9 \Omega \text{ cm}$, respectively. The cut-off wavelength of the generic PCW was located at $1,625 \text{ nm}$. The structure was then oxidized in a dry O_2 atmosphere at $1,050^\circ\text{C}$ for 20 min to increase the effective diameter of the holes in the silicon to $282 \pm 5 \text{ nm}$, and simultaneously the thickness of the top silicon layer was reduced to $200 \pm 20 \text{ nm}$. The oxidation produced glass rings with a thickness of $\sim 40 \text{ nm}$ on the inner side of the silicon holes, and the cut-off wavelength moved down to $1,552.5 \text{ nm}$. A SiO_2 spacing layer with a thickness of $\sim 1.2 \mu\text{m}$ partially filling the air holes inside the glass rings was deposited as top cladding, using a standard PECVD technique. The structure was then annealed at $1,050^\circ\text{C}$ in a dry N_2 atmosphere for 4 h. By also depositing SiO_2 on a blank test wafer, the stress of the spacing layer was measured to be $\sim 0.3 \text{ GPa}$ tensile. The deposition of SiO_2 lowered the vertical index step of the PCW to the difference between silicon ($n = 3.476$) and glass ($n = 1.445$), hence moving the cut-off wavelength up to its final value at $1,562.5 \text{ nm}$. After measuring the nonlinearity $\chi_{\text{enh},zzx}^{(2)}$ without the straining layer, a ~ 0.75 - μm -thick Si_3N_4 straining layer was deposited on the same structure using an appropriate PECVD recipe²⁶. By depositing Si_3N_4 on a test wafer with $\sim 1 \mu\text{m}$ thermal oxide on top, the stress was measured to be $\sim 1 \text{ GPa}$ compressive. Finally, the nonlinearity $\chi_{\text{enh},zzx}^{(2)}$ was measured again.

Measurement of $\chi_{\text{enh}}^{(2)}$. A fibre-based Mach-Zehnder interferometer was used to measure the $\chi_{\text{enh},zzx}^{(2)}$ value. The PCW was placed in one arm of the Mach-Zehnder interferometer and a reference LiNbO_3 phase modulator was placed in the other arm. An equipment-limited ($\sim 110 \text{ V}$ r.m.s. at 200 kHz) a.c. voltage was applied across the PCW, and the induced phase oscillation was then balanced with an identical phase oscillation from the LiNbO_3 modulator. When the modulations were balanced, that is, they had equal amplitudes, the output of the Mach-Zehnder interferometer was unaffected by the modulators. The phase oscillation and, hence, $\chi_{\text{enh},zzx}^{(2)}$ for the silicon PCW was then determined by measuring the a.c. voltage applied to the reference LiNbO_3 modulator.

Received 4 October 2005; accepted 2 March 2006.

1. Trinh, P. D., Yegnanarayanan, S., Coppinger, F. & Jalali, B. Silicon-on-insulator (SOI) phased-array wavelength multi/demultiplexer with extremely low-polarization sensitivity. *IEEE Photon. Technol. Lett.* **9**, 940–942 (1997).
2. Pavesi, L. & Lockwood, D. J. *Silicon Photonics* (Springer, Berlin, 2004).
3. Reed, G. T. & Png, C. E. J. Silicon optical modulators. *Mater. Today* **8**, 40–50 (2005).
4. Rong, H. et al. A continuous-wave Raman silicon laser. *Nature* **433**, 725–727 (2005).
5. Gibbs, W. W. A split at the core. *Sci. Am.* **291**, 96–100 (2004).
6. Madou, M. J. *Fundamentals of Microfabrication* 299–301 (CRC Press, Boca Raton, Florida, 2002).
7. Li, G. L. & Yu, P. K. L. Optical intensity modulators for digital and analog applications. *J. Lightwave Technol.* **21**, 2010–2030 (2003).
8. Butcher, P. N. & Cotter, D. *The Elements of Nonlinear Optics 5* (Cambridge Univ. Press, Cambridge, UK, 1990).
9. Soljacic, M. & Joannopoulos, J. D. Enhancement of nonlinear effects using photonic crystals. *Nature Mater.* **3**, 211–219 (2004).
10. Notomi, M. et al. Extremely large group-velocity dispersion of line-defect waveguides in photonic crystal slabs. *Phys. Rev. Lett.* **87**, 253902 (2001).
11. Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals: putting a new twist on light. *Nature* **386**, 143–149 (1997).

12. Krauss, T. F., De La Rue, R. M. & Brand, S. Two-dimensional photonic-bandgap structures operating at near-infrared wavelengths. *Nature* **383**, 699–702 (1996).
13. Joannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals: Molding the Flow of Light* (Princeton Univ. Press, Princeton, 1995).
14. Jacobsen, R. S. *et al.* Direct experimental and numerical determination of extremely high group indices in photonic crystal waveguides. *Opt. Express* **13**, 7861–7871 (2005).
15. Hitoshi, N. *et al.* Ultra-fast photonic crystal/quantum dot all-optical switch for future photonic networks. *Opt. Express* **12**, 6606–6614 (2004).
16. Soref, R. A. Silicon-based optoelectronics. *Proc. IEEE* **81**, 1687–1706 (1993).
17. Myers, R. A., Mukherjee, N. & Brueck, S. R. J. Large second-order nonlinearity in poled fused silica. *Opt. Lett.* **16**, 1732–1734 (1991).
18. Vlasov, Y. A. & McNab, S. J. Losses in single-mode silicon-on-insulator strip waveguides and bends. *Opt. Express* **12**, 1622–1631 (2004).
19. Kondo, J. *et al.* 40-Gb/s X-cut LiNbO₃ optical modulator with two-step back-slot structure. *J. Lightwave Technol.* **20**, 2110–2114 (2002).
20. Liao, W. J. *et al.* Proton-exchanged optical waveguides fabricated by glutaric acid. *Opt. Laser Technol.* **36**, 603–606 (2004).
21. Lee, K. K., Lim, D. R., Kimerling, L. C., Shin, J. & Cerrina, F. Fabrication of ultralow-loss Si/SiO₂ waveguides by roughness reduction. *Opt. Lett.* **26**, 1888–1890 (2001).
22. Soref, R. A. & Bennett, B. R. Electrooptical effects in silicon. *IEEE J. Quant. Electron.* **QE-23**, 123–129 (1987).
23. Liu, A. *et al.* A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor. *Nature* **427**, 615–618 (2004).
24. Liao, L. *et al.* High speed silicon Mach-Zehnder modulator. *Opt. Express* **13**, 3129–3135 (2005).
25. Bogaerts, W. *et al.* Nanophotonic waveguides in silicon-on-insulator fabricated with CMOS technology. *J. Lightwave Technol.* **23**, 401–412 (2005).
26. Rasmussen, F. E. *Electrical Interconnections through CMOS Wafers*. PhD. thesis, Tech. Univ. Denmark (2003); (<http://www.mic.dtu.dk/upload/institutter/mic/forskning/mems/report-31102003.pdf>).

Acknowledgements We thank R. Kjær and M. Svalgaard for their contributions. This work was supported in part by the NKT academy, the Danish Research Council for Technology and Production Sciences via the PIPE project, by NEDO via the Industrial Technology Research Area and by CINF via the Danish National Research Foundation. All generic SOI PCWs were fabricated within the framework of the European IST project PICCO and in this connection we especially thank W. Bogaerts and R. Baets.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.S.J. (Rune@com.dtu.dk).

Discovery of a 25-cm asteroid clast in the giant Morokweng impact crater, South Africa

W. D. Maier^{1,2}, M. A. G. Andreoli^{3,4}, I. McDonald⁵, M. D. Higgins¹, A. J. Boyce⁶, A. Shukolyukov⁷, G. W. Lugmair⁷, L. D. Ashwal⁴, P. Gräser², E. M. Ripley⁸ & R. J. Hart⁹

Meteorites provide a sample of Solar System bodies and so constrain the types of objects that have collided with Earth over time. Meteorites analysed to date, however, are unlikely to be representative of the entire population and it is also possible that changes in their nature have occurred with time¹. Large objects are widely believed to be completely melted or vaporized during high-angle impact with the Earth^{2,3}. Consequently, identification of large impactors relies on indirect chemical tracers, notably the platinum-group elements⁴. Here we report the discovery of a large (25-cm), unaltered, fossil meteorite, and several smaller fragments within the impact melt of the giant (>70 km diameter), 145-Myr-old Morokweng crater, South Africa. The large fragment (clast) resembles an LL6 chondrite breccia, but contains anomalously iron-rich silicates, Fe-Ni sulphides, and no troilite or metal. It has chondritic chromium isotope ratios and identical platinum-group element ratios to the bulk impact melt. These features allow the unambiguous characterization of an impactor at a large crater. Furthermore, the unusual composition of the meteorite suggests that the Morokweng asteroid incorporated part of the LL chondrite parent body not represented by objects at present reaching the Earth.

Understanding the types of extraterrestrial projectiles (asteroids and comets) that have struck Earth and formed the largest impact craters is limited by the nature of the evidence left after impact. For craters with diameters >4 km, remnant projectile fragments are generally not preserved and the impactor type can only be constrained using indirect geochemical tracers. Ni, Co and the platinum-group elements (PGEs) occur at much higher concentrations in meteorites than in average crust^{4,5}. The elements also occur in different ratios within different classes of meteorites and thus may be used to constrain the type of impactor⁶. However, the method assumes that the siderophile elements are not significantly fractionated during impact and that the bulk chemistries of large impactors match those established using the available population of small meteorites^{5–7}. The Cr isotope method has been developed more recently. It is based on the fact that all meteorite classes studied so far have Cr isotopic signatures that are different from those of Earth⁸ and are characteristic for particular classes of meteorites. The variation of the Cr isotopic composition is a result of the decay of the now-extinct parent radionuclide ⁵³Mn (half-life $T_{1/2} = 3.7$ Myr) into stable ⁵³Cr in early-accreted Solar System matter. The precise measurement of Cr isotopic abundances can unequivocally demonstrate the presence of cosmic materials in impact deposits and impact melt samples^{9,10} and, in many instances, will allow the identification of the type of the impactor.

Numerical simulations of the fate of 10-km-diameter stony

asteroids during hypervelocity ($>20 \text{ km s}^{-1}$) impacts with the Earth's surface illustrate why large projectiles can only be studied indirectly. Unlike small craters where projectiles may have suffered velocity reduction or break-up in the atmosphere, large craters are formed by massive objects that transit the atmosphere with minimal reduction in velocity. At high impact angles (30–90° to the horizontal), simulations predict peak shock pressures of 200–500 GPa that, when released, produce complete melting and partial vaporization of both the asteroid and terrestrial target rocks in the first few seconds after impact^{2,3}. Comparable simulations at lower impact angles (15°) predict peak shock pressures of ~ 100 GPa and a greater probability that some of the asteroid may survive as solid fragments. This has also been observed in experiments¹¹. However, the larger horizontal component of motion at low angles significantly increases the probability that solid or partially melted asteroid will be ejected downrange from the crater. As a consequence, except for near-vertical impacts, it is improbable that significant amounts of projectile would remain inside large impact craters³. Furthermore, given the predicted temperatures for both peak-shock and post-shock conditions (2,000–14,000 K; ref. 3), a stony asteroid would cease to exist in its original form and any geochemical tracer in the melt sheet is widely assumed to be contributed by melted or recondensed material^{2,4,5}.

In contrast, and as predicted by scaled impact experiments¹¹ and numerical modelling, fossil meteorites have been discovered in distal ejecta deposits^{12,13}, and also at the Eltanin marine impact site¹⁴. A highly altered 2.5-mm micrometeorite found within the interval spanning the Cretaceous–Tertiary (K/T) boundary in the central Pacific Ocean has been proposed as a projectile fragment from the 180-km Chicxulub crater¹³. The Chicxulub impact melt lacks significant projectile contamination¹⁵, consistent with an oblique impact scenario¹⁶, and the K/T micrometeorite is the only physical evidence presented to date that constrains the type of projectile associated with any large impact crater. In this paper the assumption that large craters contain only melted or recondensed projectile is challenged by the discovery of solid chondrite fragments up to boulder size (≥ 25 cm) in the impact melt of the Morokweng crater in South Africa. These constitute, to our knowledge, the first complete fossil stony meteorites found in an impact melt, rather than in distal ejecta. Morokweng is a large crater with a diameter of at least 70 km (refs 17–21) that formed at 145 Myr ago, coincident with the Jurassic–Cretaceous boundary^{17,18}. Younger sediments obscure the crater and most information comes from four boreholes (WF3, WF4, WF5 and M3). Of these, borehole M3 intersected ~ 870 m of impact melt^{21,22}.

¹Sciences de la Terre, Université du Québec à Chicoutimi, Chicoutimi, Quebec G7H 2B1, Canada. ²Department of Geology, University of Pretoria, Pretoria 0002, South Africa.

³South African Nuclear Energy Corporation, Pretoria 0001, South Africa. ⁴School of Geosciences, University of the Witwatersrand, Wits 2050, South Africa. ⁵School of Earth, Ocean & Planetary Sciences, Cardiff University, Cardiff CF10 3YE, UK. ⁶Scottish Universities Environmental Research Centre, East Kilbride G75 0QQ, UK. ⁷Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, USA. ⁸Department of Geological Sciences, Indiana University, Bloomington, Indiana 47405-7000, USA. ⁹Itshamba LABS - Gauteng, Wits 2050, South Africa.

The impact melt mineralogy and chemistry have been summarized previously^{19,20,22}. Impact melt pyroxenes are unusually rich in Ni (up to 0.25% NiO) and Cr (up to 0.35% Cr₂O₃ in orthopyroxene, 0.69% in clinopyroxene) compared to igneous pyroxenes in rocks of comparable MgO contents. Alongside the Clearwater East crater in Quebec²³, the Morokweng melt sheet contains the highest levels of meteoritic component, recorded by PGE concentrations, known from impact melts^{6,7,18,22}. Small nodules (<1 cm) composed of Ni-rich silicates, oxides and sulphides occur in the melt sheet, especially between 300 m and 400 m depth. These have exceptionally high PGE concentrations coupled with low initial ¹⁸⁷Os/¹⁸⁸Os ratios and may represent partially melted chondrite fragments²¹. Analyses of PGE ratios and Cr isotopes in the impact melt independently suggest that the projectile was an ordinary (type L or LL) chondrite asteroid^{6,7,10}.

Further analysis of the M3 core has revealed a 25-cm-wide meteorite clast at ~766 m depth. Apart from a thin (1 mm) corona containing brown alteration minerals (Fig. 1a) the clast is unaltered. Petrographic study of this clast revealed all the diagnostic features of a highly equilibrated chondrite breccia, including numerous subangular chondrite fragments (Fig. 1b), and well-preserved chondrules of porphyritic orthopyroxene, radial orthopyroxene (Fig. 1c) and barred olivine (Fig. 1d), up to 5 mm in width. Olivine and orthopyroxene are the main phases forming the chondrules and the mesostasis. Both minerals commonly show undulatory and mosaic extinction (Fig. 1e).

The matrix material has a well-developed granular texture with 120° triple junctions. Both olivine (Fo_{67–70}) and orthopyroxene (En_{69–73}) are compositionally homogenous within individual grains and throughout the meteorite. Additional phases, mainly occurring in the mesostasis, are plagioclase (mostly An_{33–50}, some to An₁₇), chromite (44–48% Cr₂O₃, 36–38% FeO, 2.7–3.1% TiO₂, 9–11% Al₂O₃, 2.2–2.9% MgO, 0.4–0.5% MnO) and apatite. The chondritic clast also contains abundant (~5 modal%) fine-grained pyrrhotite (Fe_{0.85–0.87}S) and pentlandite ((Fe,Ni)_{1.02–1.10}S), but no troilite (FeS) or metal (see Supplementary Information for olivine, orthopyroxene, plagioclase, chromite and sulphide data).

In addition to the large meteorite, smaller (subcentimetre-sized) fragments occur throughout the melt sheet, but especially between about 345 m and 400 m depth. These have a fine-grained, granular texture and consist predominantly of orthopyroxene showing undulating, mosaic extinction, and with compositions similar to that of the larger meteorite. In some cases, the meteorite fragments are surrounded by rims of fine-grained impact melt (Fig. 1f).

The bulk chemistry of the large clast is chondritic, although U, Th, La and Ce are enriched and Na, K are depleted compared to the average compositions for ordinary chondrites (Fig. 2a). The light-rare-earth elements (LREE), U and Th are concentrated in phosphates (apatite) during chondrite metamorphism²⁴ and the observed enrichment could reflect a higher than normal apatite content in the

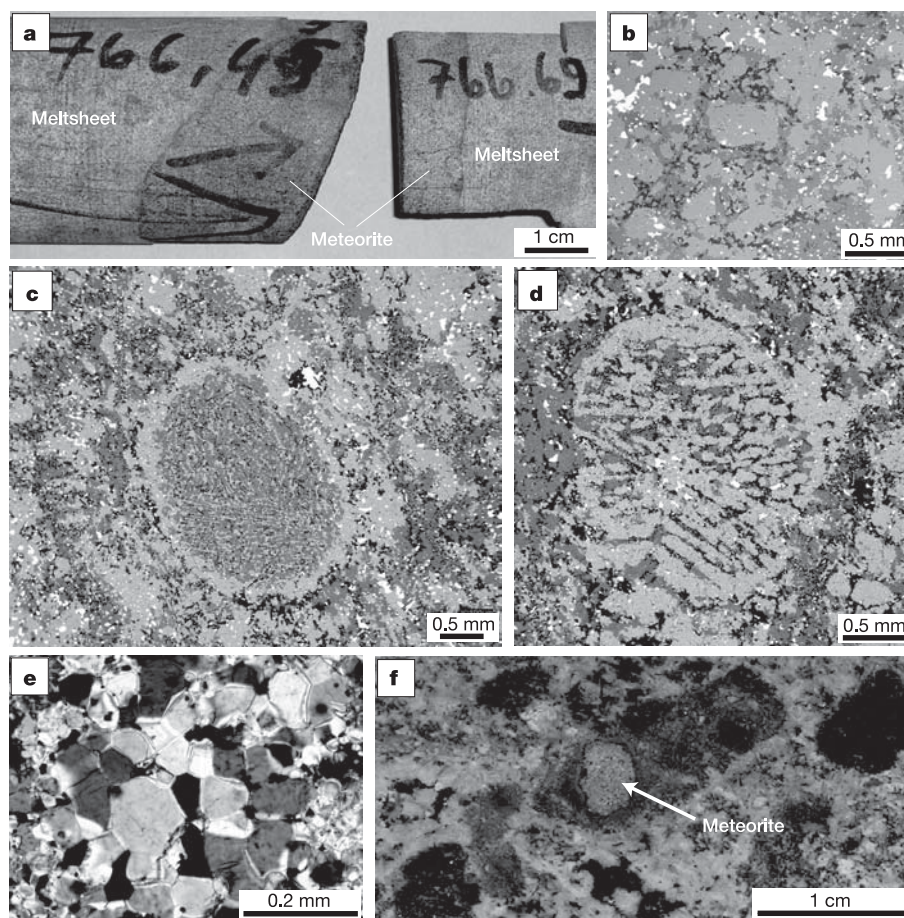


Figure 1 | Macroscopic and microscopic images of the Morokweng meteorite. **a**, Upper and lower contact of the Morokweng meteorite with the melt sheet. The meteorite is mantled by a thin corona consisting of unidentified alteration minerals, including Fe-silicates and K-rich sheet silicates. Arrow points to the bottom of the meltsheet. **b**, Fragments within the meteorite indicating that the sample is a chondrite breccia. The fragments consist largely of olivine. The matrix contains mainly

orthopyroxene (medium grey), plagioclase (dark grey) and opaques (in white), the latter consisting of sulphides and small amounts of chromite. **c**, Excentroradial, rimmed, orthopyroxene chondrule mantled by olivine. **d**, Barred olivine chondrule showing a fine-grained rim of olivine. **e**, Highly equilibrated olivines with 120° triple junctions showing mosaic textures with undulating extinction. **f**, Meteorite fragment within a clast of fine-grained norite in medium-grained leucocratic impact melt.

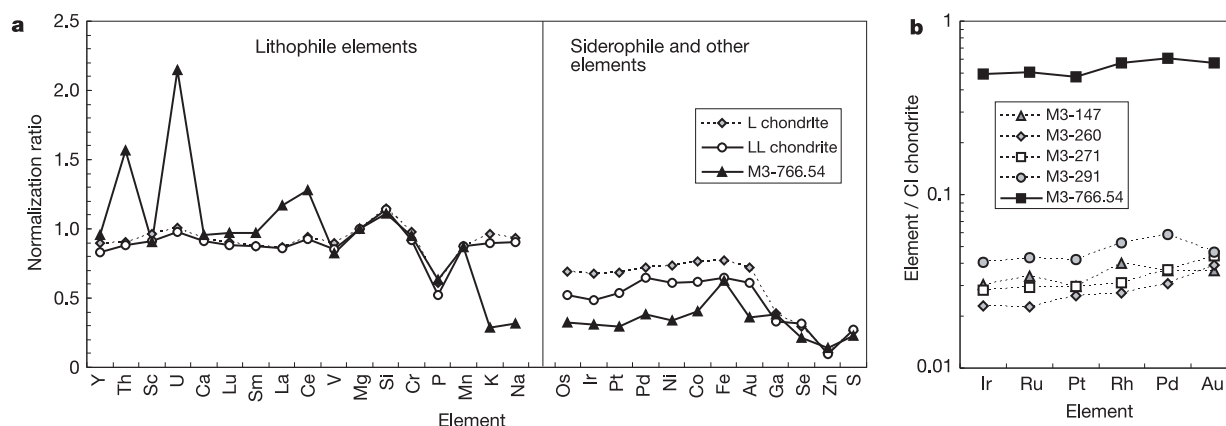


Figure 2 | Bulk chemical data for the Morokweng meteorite. **a**, Lithophile and siderophile element concentrations in the meteorite (M3-766.54) and in average L and LL chondrites normalized to Mg and CI chondrite³². **b**, PGE and Au data for the Morokweng meteorite (M3-766.54) and samples of the

Morokweng impact melt⁶ (M3-147, -260, -271 and -291), normalized to CI chondrite³³. Elements are arranged in order of decreasing 50% condensation temperature³³ (see Supplementary Table 1 for the data).

clast. Absolute PGE, Ni and Co concentrations are lower than in average LL chondrites but PGE ratios and normalized patterns of the clast are identical to those in the melt sheet^{6,7} (Fig. 2b). Despite the absence of metal, the bulk Fe content of the clast is chondritic.

Bulk S isotope analyses yielded $\delta^{34}\text{S}$ values of -0.1‰ and -0.4‰ , and individual sulphide minerals analysed by *in situ* laser combustion revealed a range of values from -0.9‰ to $+0.1\text{‰}$. This is typical of the range found in LL chondrites²⁵. By contrast, sulphides from samples of impact melt without obvious clasts gave consistently heavier $\delta^{34}\text{S}$ values ($+1.4\text{‰}$ to $+1.7\text{‰}$). This is taken to reflect a proportionally greater input of sulphur from the target lithologies ($+2.2\text{‰}$ in basement gneisses) in these samples.

An elevated $^{53}\text{Cr}/^{52}\text{Cr}$ ratio in the large clast of $0.38 \pm 0.05\epsilon$ (ϵ is the deviation in parts in 10^4 from the terrestrial $^{53}\text{Cr}/^{52}\text{Cr}$ ratio) clearly indicates meteoritic material. This value is higher than in the Morokweng impact melt samples ($\sim 0.24\epsilon$ and $\sim 0.27\epsilon$; ref. 10) due to a larger proportion of extraterrestrial Cr. It also is inconsistent with both carbonaceous or enstatite chondrites but indicative of an ordinary chondrite⁸.

The chondritic bulk composition and the presence of fine-grained impact melt around the meteorite clasts, coupled with Cr isotope and PGE geochemistry in the large clast that matches the meteoritic signatures in the bulk melt sheet, confirm that the meteorite boulder in M3 represents a fragment of the original asteroid that struck Morokweng. This discovery has important implications. First, it could suggest that models for the evolution of the projectile during impact are currently incomplete. Even if the fragment came from the projectile's trailing edge, where shock pressures should be lowest, currently available models (which assume a 20 km s^{-1} impact velocity) still predict at least partial melting of the whole projectile in all but the most oblique scenarios^{2,3}. However the high degree of meteoritic contamination observed at Morokweng favours a high-angle impact and would appear to rule out an oblique event. Single projectile models should be re-evaluated to see whether lower impact velocities used in some recent simulations ($12\text{--}16 \text{ km s}^{-1}$; ref. 26) coupled with high impact angles can simulate the mixture of possibly melted and definitely unmelted fragments found at Morokweng, or whether alternatives such as an additional trailing projectile might be required. Second, it directly confirms the conclusions of previous studies on the Morokweng impact melt that used geochemistry alone to constrain the nature of the impactor as an L or LL ordinary chondrite^{6,7,10}. It thereby supports the assumption that the identity of a large impactor can be inferred, at least to the level of a chondrite group, using siderophile element data from small meteorites.

Nevertheless, other elements of the meteorite's mineralogy are unusual. The Fe contents of olivine and chromite, and Fe and Ca

contents of orthopyroxene are higher than most LL4–LL6 chondrites (ref. 27; Fig. 3) and more closely resemble the rare LL7 chondrites (for example, Sahara 97937; ref. 28). There is no evidence to suggest that the absence of metal and abundance of sulphide are the result of contamination or interaction with the impact melt. Instead, they appear to be primary and might be an extreme product of C–O–H–S-bearing fluids believed to have permeated highly equilibrated ordinary chondrites during parent body metamorphism²⁹. Redistribution of metallic Fe into silicates occurs with increasing metamorphism in LL5 and LL6 chondrites²⁹, and the observed fractionation between Fe and the other siderophiles may be influenced by different reaction pathways involving oxygen or sulphur. Similarly, the depletions in Na, K and Cs may also reflect removal of these elements in a metamorphic fluid. If the mineralogy was imposed by parent body metamorphism, the fragment could be derived from a previously unknown interior portion of the LL chondrite parent body that is not sampled by the objects that currently supply small meteorites to Earth. Whether this reflects a temporal change in the type and composition of impactors generally (see, for example, ref. 1), or

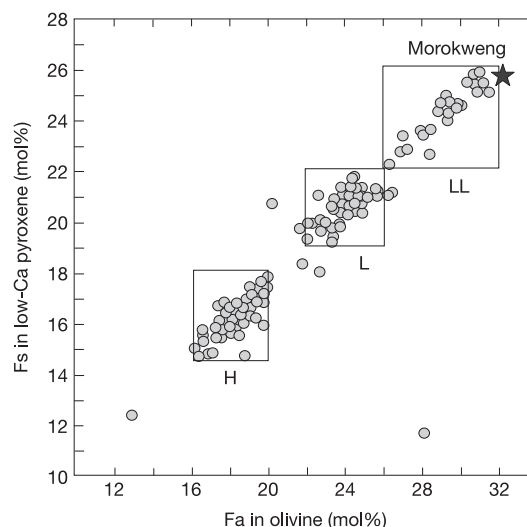


Figure 3 | Mineral compositional data for the Morokweng meteorite. Diagram shows ferrosilite (Fs) content of orthopyroxene versus fayalite (Fa) content of olivine in metamorphic grade 4–6 ordinary chondrites (figure modified after ref. 27). Morokweng olivines and orthopyroxenes (indicated by star symbol) have higher Fs and Fa contents than most other LL, L and H chondrites. See Supplementary Information for the data.

whether large impacts like Morokweng result from periodic sampling of a different asteroid population is unknown and will only become clearer if it is possible to find and characterize similar fossil meteorites in other large impact craters.

METHODS

The bulk geochemistry of the clast was characterized by inductively coupled plasma optical emission spectrometry and inductively coupled plasma mass spectrometry at Cardiff University. Sample preparation techniques and the fire assay procedures used for PGE analysis are outlined elsewhere³⁰. Sulphur contents were determined by a combustion idiometric procedure using Laboratory Equipment Company (LECO) titration and Se was determined by instrumental neutron activation analysis after preconcentration of the element in a Ni sulphide bead at the University of Quebec at Chicoutimi. Mineral compositions were determined using a CAMECA SX100 electron microprobe at the University of Pretoria. Whole rock sulphur isotopic measurements were carried out at Indiana University in Bloomington, USA, and the Scottish Universities Environmental Research Centre (SUERC). *In situ* sulphur isotopic measurements by laser combustion were carried out at the SUERC³¹. The Cr isotope measurements were conducted using the thermal ionization Micromass Sector 54 mass spectrometer at the University of California, San Diego. Sample preparation, chemical procedure, and mass spectrometry were similar to those described elsewhere⁸.

Received 20 January; accepted 24 March 2006.

1. Benoit, P. H. & Sears, D. W. G. The breakup of a meteorite parent body and the delivery of meteorites to Earth. *Science* **255**, 1685–1687 (1992).
2. Pierazzo, E. & Melosh, H. J. Hydrocode modelling of Chicxulub as an oblique impact event. *Earth Planet. Sci. Lett.* **165**, 163–176 (1999).
3. Pierazzo, E. & Melosh, H. J. Hydrocode modelling of oblique impacts: The fate of the projectile. *Meteorit. Planet. Sci.* **35**, 117–130 (2000).
4. Koeberl, C. in *Meteorites: Flux with Time and Impact Effects* (eds Grady, M. M., Hutchinson, R., McCall, G. J. H. & Rothery, R. A.) 133–153 (The Geological Society, London, 1998).
5. Grieve, R. A. F. Terrestrial impact: the record in the rocks. *Meteoritics* **26**, 175–194 (1991).
6. McDonald, I., Andreoli, M. A. G., Hart, R. J. & Tredoux, M. Platinum-group elements in the Morokweng impact structure, South Africa: evidence for the impact of a large ordinary chondrite projectile at the Jurassic-Cretaceous boundary. *Geochim. Cosmochim. Acta* **65**, 113–123 (2001).
7. Koeberl, C. & Reimold, W. U. Geochemistry and petrography of impact breccias and target rocks from the 145 Ma Morokweng impact structure, South Africa. *Geochim. Cosmochim. Acta* **67**, 1837–1862 (2003).
8. Lugmair, G. W. & Shukolyukov, A. Early solar system timescales according to ⁵³Mn–⁵³Cr isotope systematics. *Geochim. Cosmochim. Acta* **62**, 2863–2886 (1998).
9. Shukolyukov, A. & Lugmair, G. W. Isotopic evidence for the Cretaceous-Tertiary boundary impactor and its type. *Science* **282**, 927–929 (1998).
10. Shukolyukov, A., Lugmair, G. W., Koeberl, C. & Reimold, W. U. Chromium in the Morokweng impact melt: isotopic evidence for extraterrestrial components and type of impactor. *Meteorit. Planet. Sci.* **34**, A107–A108 (1999).
11. Schultz, P. H. & Sugita, S. Fate of the Chicxulub impactor. *Lunar Planet. Sci.* **XXVIII**, 1261–1262 (1997).
12. Thorslund, P. & Wickman, F. E. Middle Ordovician chondrite in fossiliferous limestone from Brunflo, central Sweden. *Nature* **289**, 285–286 (1981).
13. Kyte, F. T. A meteorite from the Cretaceous/Tertiary boundary. *Nature* **396**, 237–239 (1998).
14. Kyte, F. T. Unmelted meteoritic debris collected from Eltanin ejecta in Polarstern cores from expedition ANT XII/4. *Deep-sea Res. II* **49**, 1063–1071 (2002).
15. Tagle, R. *et al.* Platinum-group elements in impactites of the ICDP Chicxulub drill core Yaxcopoil-1: are there traces of the projectile? *Meteorit. Planet. Sci.* **39**, 1009–1016 (2004).
16. Schultz, P. H. & D'Hondt, S. Cretaceous-Tertiary (Chicxulub) impact angle and its consequences. *Geology* **24**, 963–967 (1996).
17. Hart, R. J. *et al.* Late Jurassic age for the Morokweng impact structure, southern Africa. *Earth Planet. Sci. Lett.* **147**, 25–35 (1997).
18. Koeberl, C., Armstrong, R. A. & Reimold, W. U. Morokweng, South Africa: a large impact structure of Jurassic-Cretaceous boundary age. *Geology* **25**, 731–734 (1997).
19. Andreoli, M. A. G., Ashwal, L. D., Hart, R. J. & Huizenga, J. M. A. in *Large Meteorite Impacts and Planetary Evolution II* (eds Dressler, B. O. & Sharpton, V. L.) 91–108 (Special Paper 339, Geological Society of America, Boulder, Colorado, 1999).
20. Reimold, W. U., *et al.* in *Large Meteorite Impacts and Planetary Evolution II* (eds Dressler, B. O. & Sharpton, V. L.) 61–90 (Special Paper 339, Geological Society of America, Boulder, Colorado, 1999).
21. Hart, R. J., Cloete, M., McDonald, I., Carlson, R. W. & Andreoli, M. A. G. Siderophile-rich inclusions from the Morokweng impact melt sheet, South Africa: possible fragments of a chondritic meteorite. *Earth Planet. Sci. Lett.* **198**, 49–62 (2002).
22. Maier, W. D. *et al.* The Morokweng impact melt sheet, South Africa: a reconnaissance study with implications for Ni-Cu-PGE sulphide mineralization. *Trans. Inst. Min. Metall.* **112**, B150–B152 (2003).
23. Palme, H., Janssens, M.-J., Takahashi, H., Anders, E. & Hertogen, J. Meteoritic material at five large impact craters. *Geochim. Cosmochim. Acta* **42**, 313–323 (1978).
24. Ebihara, M., Shinotsuka, K. & Kong, P. Redistribution of trace elements during metamorphism of ordinary chondrites. *Lunar Planet. Sci.* **XXVII**, 329–330 (1996).
25. Gao, X. & Thiemens, M. H. Variations of the isotopic composition of sulfur in enstatite and ordinary chondrites. *Geochim. Cosmochim. Acta* **57**, 3171–3176 (1993).
26. Ivanov, B. A. & Stöffler, D. The Steinheim impact crater, Germany: modelling of a complex crater with central uplift. *Lunar Planet. Sci.* [CD ROM] XXXVI, 1443 (2005).
27. Brearley, A. J. & Jones, R. H. in *Planetary Materials* (ed. Papike, J. J.) 3–1–3–398 (Vol. 36, Reviews in Mineralogy, Min. Soc. America, Washington DC, 1996).
28. Grossman, J. N. The Meteoritical Bulletin, No 82, 1998 July. *Meteorit. Planet. Sci.* **33**, A221–A239 (1998).
29. McSween, H. Y. & Labotka, T. C. Oxidation during metamorphism of the ordinary chondrites. *Geochim. Cosmochim. Acta* **57**, 1105–1114 (1993).
30. McDonald, I., Irvine, G. J., de Vos, E., Gale, A. S. & Reimold, W. U. Geochemical search for impact signatures in possible impact-generated units associated with the Jurassic-Cretaceous boundary in southern England and northern France. in *Biological Processes Associated with Impacts* (eds Cockell, C., Gilmour, I. & Koeberl, C.) 257–286 (Springer Verlag, Berlin, 2006).
31. Wagner, T., Boyce, A. J. & Fallick, A. E. Laser combustion analysis of $\delta^{34}\text{S}$ of sulfosalt minerals: determination of the fractionation systematics and some crystal-chemical considerations. *Geochim. Cosmochim. Acta* **66**, 2855–2863 (2002).
32. Wasson, J. T. & Kallemeyn, G. W. Composition of chondrites. *Phil. Trans. R. Soc. Lond. A* **325**, 535–544 (1988).
33. Lodders, K. Solar system abundances and condensation temperatures of the elements. *Astrophys. J.* **591**, 1220–1247 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was logistically supported by Business Venture Investments No. 33 (Pty) Ltd. (M.A.G.A.). Financial support was provided by S.-J. Barnes, Canada Research Chair in Magmatic Ore Deposits, UQAC. In addition, analytical grants were provided by the University of Pretoria (W.D.M.), and by NASA (A.S. and G.W.L.). ICP analyses at Cardiff University were supported by the Leverhulme Trust (I.McD.) and by a JIF award from the NERC. A.J.B. is supported by NERC Services and Facilities funding to the Isotope Community Support Facility at SUERC. We thank H. Gregoire for assistance in producing the EMS images, and R. Ash and R. Burgess for comments on an early version of the manuscript.

Author Contributions All co-authors contributed in the form of data generation and/or discussion and critical comment.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.D.M. (wolfgang.maier@uqac.ca).

New carbon dates link climatic change with human colonization and Pleistocene extinctions

R. Dale Guthrie¹

Drastic ecological restructuring, species redistribution and extinctions mark the Pleistocene–Holocene transition, but an insufficiency of numbers of well-dated large mammal fossils from this transition have impeded progress in understanding the various causative links¹. Here I add many new radiocarbon dates to those already published on late Pleistocene fossils from Alaska and the Yukon Territory (AK–YT) and show previously unrecognized patterns. Species that survived the Pleistocene, for example, bison (*Bison priscus*, which evolved into *Bison bison*), wapiti (*Cervus canadensis*) and, to a smaller degree, moose (*Alces alces*), began to increase in numbers and continued to do so before and during human colonization and before the regional extinction of horse (*Equus ferus*) and mammoth (*Mammuthus primigenius*). These patterns allow us to reject, at least in AK–YT, some hypotheses of late Pleistocene extinction: ‘Blitzkrieg’ version of simultaneous human overkill², ‘keystone’ removal³, and ‘palaeo-disease’⁴. Hypotheses of a subtler human impact and/or ecological replacement or displacement are more consistent with the data. The new patterns of dates indicate a radical ecological sorting during a uniquely forage-rich transitional period, affecting all large mammals, including humans.

The gradient from Pleistocene to Holocene covered many millennia, but I show here that in AK–YT the portion of the Pleistocene–Holocene transition (P/HT) most critical for large mammals and their forage occurred from 13,500 to 11,500 radiocarbon years before present (13.5–11.5 kyr BP). The present study focuses on six large mammal species, two that became extinct at the P/HT (mammoth and horse) and four that did not (bison, wapiti, moose and humans). Interior AK–YT is unique in North America for such a study, not only because of its location as a connecting link between the hemispheres but also because many plant and animal species encounter the outer limits of their ecological tolerance at these latitudes, thus often magnifying evidence of climatic–biotic shifts. Fortunately, remains of Pleistocene large mammals are abundant in this region and are exceptionally preserved by permafrost, retaining most of their easy-to-date collagen; indeed, many specimens retain marrow and connective tissue⁵.

Although many published dates for fossils of mammoth and horse already existed, these species were so important that I concentrated on them, more than doubling the number of previously published dates for these two species in AK–YT. New dates for moose and wapiti were obtained because of the previous rarity of radiocarbon dates for these species in AK–YT. I did not add to the pool of bison dates because it was well developed by previous work by myself and others^{5–7}. For information about humans, I used the accumulated pool of dates from Alaskan archaeological sites⁸. The resulting patterns are illustrated in Fig. 1. Details and dates are provided in Supplementary Information.

An explanation for the unexpected radiocarbon patterns becomes

clearer when they are placed in the broader palaeoecological setting (Fig. 2). Fortunately, the character and sequence of the P/HT vegetational changes, lake formation and soil paludification in interior AK–YT have been studied by means of dozens of pollen cores and a wealth of macrofossils^{9–11}.

In terms of how these date patterns affect extinction theories, neither synchronous extinctions nor demographic morbidity gaps as predicted by the virulent multi-species ‘palaeo-disease’ hypothesis⁴ are apparent in this chronology. Nor do these dates lend support to the ‘keystone’ hypothesis³, which proposes that the removal of

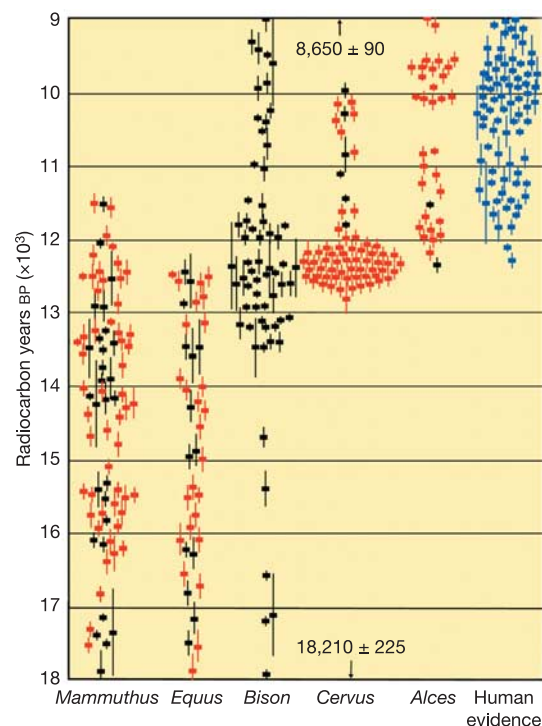


Figure 1 | A plot of radiocarbon dates on AK–YT mammoth, horse, bison, wapiti and moose bones that fall within the 18–9 kyr BP interval, shown alongside dated archaeological material (hearth charcoal mainly) from the same regions⁸. See Supplementary Information for further discussion. Error bars show s.d. Dates from this study are shown in red; other published faunal dates are shown in black, and archaeological dates are blue. The two near-outlier dates listed for wapiti indicate their presence just off the chart. Patterns in this graph are assumed to be a rough indicator of relative abundance and changes with time within each species. A discussion on specimen selection, actual dates and their detailed provenance for all species is provided in Supplementary Information.

¹Institute of Arctic Biology, University of Alaska, Fairbanks, Alaska 99709, USA.

mammoths by humans led to vegetational transformation, which in turn led to a secondary set of extinctions. In this AK–YT chronology, horse extinction seems to have preceded the demise of the proposed keystone species, woolly mammoth, by a calendar millennium¹².

The relevance of this study to the human overkill hypothesis is more complicated. Currently, archaeological visibility of humans in Alaska began soon before 12 kyr BP (Supplementary Information), which just abuts the final dates for horse fossils (Fig. 1). The millennium-wide disjunction of first human dates and terminal mammoth dates indicates that humans could well have had a hand in the gradual extinction of mammoths, but not as in a century-scale ‘Blitzkrieg’ overkill² in which a newly arriving wave of super-efficient human hunters broadly and abruptly devastated local megafaunas. There are two archaeological sites from the period of human–mammoth overlap that contain tusk fragments or tools of ivory^{13,14}, although that does not necessarily indicate mammoth hunting, much less imply overkill.

Unlike the fossil record in Europe, there is as yet no evidence of humans hunting horses in AK–YT, nor is there in adjacent Siberia. In contrast, there is good evidence from early Alaskan archaeological sites that bison and wapiti were successfully hunted^{13–15}. The problem of horse extinction in AK–YT is further compounded by the observation of a decline in horse body size before both human arrival and horse extinction¹². If humans had a role in horse extinction it occurred on top of other major ecological change that was already well underway. In fact, many species seem to have become extinct in AK–YT well before the period covered by this chart. No AK–YT fossils (see the US–Canadian radiocarbon website in Supplementary Information) of the stilt-legged equids (*Equus* sp.), stag moose (*Cervalces*), camels (*Camelops*), giant beaver (*Castoroides*), ground sloth (*Megalonyx*), extinct musk oxen (*Praeovibos*), mastodon

(*Mammut*), short-faced bear (*Arctodus*) and scimitar cat (*Homo-therium*) have been recorded after the Last Glacial Maximum (LGM), about 18 kyr BP.

Why were the two grazing specialists, bison and wapiti, absent or rare in AK–YT for five millennia before the transitional period, and how are we to understand their later remarkable boom? We know that conditions in this part of the mammoth steppe during the LGM were extreme. Dune fields, thick loess deposits and the virtual absence of lake sediments indicate arid and windy conditions with a largely treeless, short-grass–sedge–sage sward^{5,16}. These conditions of sparse forage and lack of riparian cover and alternate forage might account for the gap in bison and wapiti dates, but I suspect that the competitive advantage of the two caecalid grazing specialists, mammoth and horse, also had a role.

A large hindgut diverticulum, the caecum, gives today’s elephants, horses, rhinos and other caecalids the potential for a rapid gut throughput, allowing them to tolerate high volumes of poor-quality forage⁵, although at the cost of more conservative growth and reproduction. Mammoths and horses (and woolly rhinos, *Coleodonta*, in adjacent Siberia) were thus better adapted to quantities of exposure-leached winter graminoids that characterized the mammoth steppe. In addition, winter range is always the bottleneck for northern herbivores¹⁷. Note that if caecalids are removed from the list of extinct species across northern Eurasia and Beringia, extinctions of large herbivores at the end of the Pleistocene become a minor event.

The first signal of the P/HT appears in AK–YT as a pollen shift arising just before 13.5 kyr BP^{9–11,18}. Bison, apparently local although not so abundant, judging from Fig. 1, was the first species to expand at the beginning of the P/HT. Because wapiti fossil dates are completely missing from the 5 kyr before their expansion, we might assume that it reflects their local absence. Lag in wapiti expansion might simply be related to the dynamics of colonization distance. It is possible that wapiti, less dependent than moose on a mesic habitat, colonized from arid northern Eurasia¹⁹, closer to AK–YT. The lateness of AK–YT radiocarbon dates for humans and for moose (an obligate browser), as well as the present pattern of no early dates for humans and moose in adjacent Siberia, suggests that both species might have been colonizing at the same time from more distant mid-continental Asia.

Water-limited cold steppe vegetation apparently responded early in the P/HT to the irregular shift towards slightly more moisture and warmer summer temperatures with an expansion of meadowland graminoids and forbs. Tree willow (*Salix*) also became more abundant^{20,21}. Lake histories and pollen cores^{9,11,22} across the north show that the trend of increasing temperature and moisture ultimately pushed the adaptive vegetational balance firmly towards the now familiar Holocene lakes, bogs, shrub tundra, forests and low-nutrient acidic soils. This resulted in communities of conservative plants highly defended against herbivory²³ and supporting a small biomass of large mammals³. Dwarf birch (*Betula*) (see Fig. 2) is especially toxic²³. Pivotal species shifts are echoed in the fossil invertebrate record as well¹⁸.

Even though willow is primarily entomophilous, the influx and percentage of willow pollen were often greater during this transitional episode than before or afterwards, ranging up to 20% of the pollen spectra^{9,11}. Tree willow macrofossils also appear during this period. The digestible energy content and nutrients of early-growth willow leaves are higher than those of virtually all other northern plants^{17,23}. Even though tree willows currently occur at only modest densities, their leaves and stems are a critical part of AK–YT moose diets¹⁷. Indeed, captive and reintroduced wapiti and bison (which are grazers, but eclectic ones) make heavy use of willows¹⁷.

Despite regional differences, this model of linkage between climate, extinctions and human colonization is generally consistent with data from northern Eurasia¹⁸. Although the broader question of human impact on large mammals in AK–YT, as in Eurasia, remains somewhat open and must be debated in the context of underlying

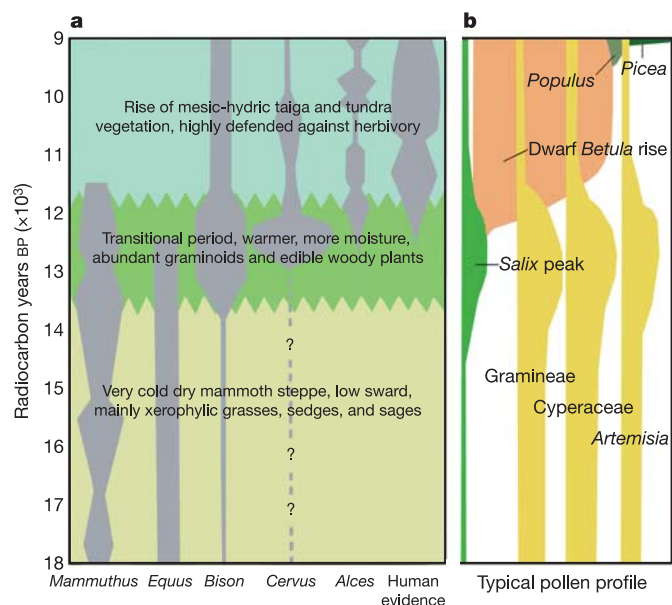


Figure 2 | A visual model linking large mammal date patterns (Fig. 1) to a changing ecological context. a, Previous interpretations of the climate–vegetation events are overlain with ‘ghost forms’ of the large mammal chronology patterns from Fig. 1. The central green band highlights the episode that I refer to as the ‘transitional period’. **b**, Generalized pollen influx profiles from several pollen studies in interior AK–YT, using key plant groups^{9–11} with emphasis on sites that used the more precise AMS core dating (because conventional radiometric dating of core grab samples generally overestimates time⁹). Patterns from individual pollen cores vary somewhat, yet all have common features; I have smoothed and stressed those regular characters in this generalized graph.

ecological trends^{24,25}, it has become increasingly clear that large mammals had a significant impact on humans. Warmer and wetter summers²⁶, an increasing availability of wood⁹, and technological innovation²⁷ each probably had a role in the human colonization of AK–YT, but archaeological refuse clearly illustrates the crucial role of large mammal (at least bison and wapiti) resources as well as the increasing numbers of migratory waterfowl and salmonids in the Holocene¹⁴. It cannot be overemphasized that vertebrates, especially large mammals, have historically formed the central resource of northern peoples²⁸. These new data indicate that humans might have been not so much riding down the demise of the Pleistocene mammoth steppe as they were being carried into AK–YT on a unique tide of resource abundance created by the P/HT.

METHODS

The collection of specimens for dating, dating potential and an evaluation of the possibility of biases are discussed in Supplementary Information, along with museum acquisition data, site location and provenance. In addition to those listed, I also included for dating six samples of mastodon (*Mammuth americanus*) and five of stag moose (*Alces latifrons*, which evolved into *Cervalces latifrons*), in case these species were part of the above story, but all 11 samples proved to be infinite or at least at the far outer radiocarbon range. A total of 23 samples of a stilt-legged²⁹ AK–YT *Equus* species were also submitted for dating; all of them dated to before 30 kyr BP¹². I chose woolly mammoths and cabaloid horses as icons of the open-ground faunas of the mammoth steppe. Moose, a browser, was chosen to represent the Holocene fauna, as this species is particularly well adapted to mesic habitats within boreal forests¹⁷. The earliest date for moose came from ref. 30. Wapiti and bison were used because preliminary evidence suggested that their fossils were common in Holocene sediments¹⁵ and present in early Holocene archaeological sites^{14,15}. Nomenclatural designations are rather straightforward except for bison⁷, which was evolving rapidly from the larger *Bison priscus* into the smaller *Bison bison* during the latest Pleistocene. All my samples (red in Fig. 1) were run at the US National Accelerator Mass Spectrometer (AMS) Radiocarbon Laboratory at Tucson, Arizona, using their conventional pretreatment¹².

Received 2 September 2005; accepted 25 January 2006.

- Grayson, D. K. & Meltzer, D. J. A requiem for North American overkill. *J. Archaeol. Sci.* **30**, 585–593 (2003).
- Martin, P. S. in *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 345–403 (Univ. of Arizona Press, Tucson, 1984).
- Zimov, S. A. et al. Steppe–tundra transition: a herbivore-driven biome shift at the end of the Pleistocene. *Am. Nat.* **146**, 765–794 (1995).
- MacPhee, R. D. E. et al. Radiocarbon chronologies and extinction dynamics of the Late Quaternary mammalian megafauna of the Taimyr Peninsula, Russian Federation. *J. Archaeol. Sci.* **29**, 1017–1042 (2002).
- Guthrie, R. D. *Frozen Fauna of the Mammoth Steppe* (Univ. of Chicago Press, Chicago, 1990).
- Stephenson, R. O., Gerlach, S. C., Guthrie, R. D. & Mills, R. in *Wildlife and People in Northern North America. Essays in Honor of R. Dale Guthrie* (eds Gerlach, S. C. & Murray, M. S.) 125–159 (Br. Archaeol. Rep. Int. Ser., London, 2001).
- Shapiro, B. et al. Rise and fall of Beringian steppe bison. *Science* **306**, 1561–1565 (2004).
- Potter, B. *Site Structure and Organization in Central Alaska: Archaeological Investigations at Gerstle River* Thesis, Univ. Alaska, Fairbanks (2004).
- Bigelow, N. H. & Powers, W. R. Climate, vegetation, and archaeology 14,000–9,000 cal yr B.P. in Central Alaska. *Arctic Anthropol.* **38**, 171–195 (2001).
- Ager, T. A. in *Late Quaternary Environments of the United States* Vol. 2 (ed. Wright, H. E. Jr) 128–141 (Univ. Minnesota Press, Minneapolis, 1983).
- Anderson, P. M., Bartlein, P. J., Brubaker, L. B., Gajewski, K. & Ritchie, J. C. Vegetation–pollen–climate relationships for arcto-boreal region of North America and Greenland. *J. Biogeogr.* **18**, 565–582 (1991).
- Guthrie, R. D. Rapid size decline in Alaskan Pleistocene horse before extinction. *Nature* **426**, 169–171 (2003).
- Bever, M. R. & Kunz, M. L. (eds) Between two worlds: Late Pleistocene cultural and technological diversity in eastern Beringia. *Arctic Anthropology* Vol. 38 (Univ. of Wisconsin Press, Madison, 2001).
- Yesner, D. R. Human dispersal into interior Alaska: antecedent conditions, mode of colonization, and adaptations. *Quat. Sci. Rev.* **20**, 315–327 (2001).
- Powers, W. R. W., Guthrie, R. D. & Hoffecker, J. F. Dry Creek: *Archaeology and Paleocology of a Late Pleistocene Alaskan Hunting Camp* (Dry Creek Rep. to the Natl Park Serv., Fairbanks, Alaska, 1983).
- Goetcheus, V. G. & Birks, H. H. Full-glacial upland tundra vegetation preserved under tephra in Beringia National Park, Seward Peninsula, Alaska. *Quat. Sci. Rev.* **20**, 135–147 (2001).
- Franzmann, A. W. & Swartz, C. C. *Ecology and Management of the North American Moose* (Smithsonian Inst. Press, Washington DC, 2002).
- Sher, A., Kuzmina, S. A., Kuznetsova, T. V. & Sulerzhitsky, L. D. New insights into the Weichselian Environment and climate of the east Siberian Arctic, derived from fossil insects, plants, and mammals. *Quat. Sci. Rev.* **24**, 533–569 (2005).
- Boeskorov, G. G. in *Univ. of Alaska AQC Workshop Publ. 3rd Int. Mammoth Pre-Conference 8* (Alaska Quaternary Center Press, Fairbanks, 2003).
- Hopkins, D. M., Smith, P. A. & Matthews, J. V. Jr. Dated wood from Alaska and the Yukon—implications for forest refugia in Beringia. *Quat. Res.* **15**, 217–249 (1981).
- Porter, S. C., Pierce, K. L. & Hamilton, T. D. in *Late Quaternary Environments of the United States* Vol. 1 (ed. Wright, H. E. Jr) 71–114 (Univ. Minnesota Press, Minneapolis, 1983).
- Lozhkin, A. V. et al. Late Quaternary lacustrine pollen records from southwestern Beringia. *Quat. Res.* **39**, 314–324 (1993).
- Bryant, J. P. & Kuropat, P. J. Selection of winter forage by subarctic browsing vertebrates: the role of plant chemistry. *Annu. Rev. Ecol. Syst.* **11**, 261–285 (1980).
- Barnosky, A. D. et al. Assessing the causes of late Pleistocene extinctions on the continents. *Science* **306**, 70–75 (2004).
- Graham, R. W. & Lundelius, E. L. Jr. in *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 223–250 (Univ. of Arizona Press, Tucson, 1984).
- Kutzbach, J. E. & Street-Perrot, F. A. Milankovitch forcing of fluctuations in the level of tropical lakes from 18 to 0 kyr BP. *Nature* **317**, 130–134 (1985).
- Slobodkin, S. Western Beringia at the end of the Ice Age. *Arctic Anthropol.* **38**, 31–47 (2001).
- Dumond, E. E. & Bland, R. L. Holocene prehistory of the northernmost North Pacific. *J. World Prehist.* **9**, 401–451 (1995).
- Weinstock, J. et al. Evolution, systematics, and phylogeography of Pleistocene horses in the New World. *PLoS Biol.* **3**, e241 (2005).
- Matheus, P. M. *Chronology and Ecology of a Late Quaternary Large Mammal Assemblage from Northern Alaska* (US Dept Interior Rep. LBP 0119, Bur. Land Mgmt Northern Field Office, Fairbanks, Alaska, 2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank curators R. Tedford, R. Gangloff, R. Harington and R. Purdy for dating access to museum collections, and M. L. Guthrie, F. Dean and D. Murray for suggestions that improved the manuscript. I also thank the NSF and the Tucson AMS laboratory.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to R.D.G. (frrdg@uaf.edu).

LETTERS

Sympatric speciation in palms on an oceanic island

Vincent Savolainen¹, Marie-Charlotte Anstett², Christian Lexer¹, Ian Hutton³, James J. Clarkson¹, Maria V. Norup^{1,4}, Martyn P. Powell¹, David Springate¹, Nicolas Salamin⁵ & William J. Baker¹

The origin of species diversity has challenged biologists for over two centuries. Allopatric speciation, the divergence of species resulting from geographical isolation, is well documented¹. However, sympatric speciation, divergence without geographical isolation, is highly controversial². Claims of sympatric speciation must demonstrate species sympatry, sister relationships, reproductive isolation, and that an earlier allopatric phase is highly unlikely¹. Here we provide clear support for sympatric speciation in a case study of two species of palm (Arecaceae) on an oceanic island. A large dated phylogenetic tree shows that the two species of *Howea*, endemic to the remote Lord Howe Island, are sister taxa and diverged from each other well after the island was formed 6.9 million years ago³. During fieldwork, we found a substantial disjunction in flowering time that is correlated with soil preference. In addition, a genome scan^{4,5} indicates that few genetic loci are more divergent between the two species than expected under neutrality, a finding consistent with models of sympatric speciation involving disruptive/divergent selection². This case study of sympatric speciation in plants provides an opportunity for refining theoretical models on the origin of species, and new impetus for exploring putative plant and animal examples on oceanic islands.

Speciation, the division of populations into evolutionarily independent units, involves genetic separation and phenotypic differentiation. Genetic divergence following geographic isolation gives rise to allopatric speciation: “the conceptual rationale is simply that, given enough time, speciation is an inevitable consequence of populations evolving in allopatry”⁶. Numerous empirical examples support this uncontroversial scenario¹. In theory, however, populations can become genetically isolated without geographical separation, resulting in sympatric speciation, a much more contentious model. Sympatric speciation is more strictly defined as the emergence of two species from a population in which mating has been random with respect to the place of birth of the mating partners².

Mathematical models have shown that sympatric speciation is possible^{2,7–10}, but very few examples have been documented in nature^{11,12}. Cichlid fish seem to have radiated sympatrically in African crater lakes. Molecular phylogenetic analyses show that the fish species in each lake share a common ancestor, with sexual selection and ecology possibly driving speciation¹³. Second, races of apple and hawthorn maggot have shifted to different hosts in sympatry and differ in reproductive behaviour and breeding time¹⁴. Third, a genetic study of African indigobirds, which are host-specific brood parasites, showed that they might have recently speciated sympatrically after new hosts were colonized¹⁵. These examples are all from animal taxa with relatively large continental geographic distributions. This leaves the door open to controversy, given that truly convincing cases of sympatric speciation must involve biogeographic and phylogenetic histories that make the existence of an allopatric phase highly unlikely¹. For this reason,

we focused on a plant species-pair confined to a remote oceanic island.

Lord Howe Island (Fig. 1a) is a small subtropical island of less than 12 km², situated 580 km off the eastern coast of Australia. The island was formed by volcanic activity 6.4–6.9 million years (Myr) ago³. Lord Howe Island is the most southerly member of a 1,000-km chain of nine underwater volcanoes along the Lord Howe Rise. The closest link in the Lord Howe Island chain is Elizabeth Reef, 160 km to the north; this seamount was an island 10.2 Myr ago³. Lord Howe Island



Figure 1 | Lord Howe Island and its endemic palms. **a**, Lord Howe Island is a World Heritage Site, and a permanent park preserve now protects 70% of the island. The waters surrounding Lord Howe Island are protected as a marine park, which also holds the world's southernmost coral reef. The island is inhabited by approximately 300 residents, and less than 20% of the vegetation has been disturbed. **b**, The kentia or thatch palm, *Howea forsteriana*, is characterized by multiple spikes in each inflorescence and has straight leaves with drooping leaflets. **c**, The curly palm, *H. belmoreana*, bears a single spike in each inflorescence and has recurved leaves with ascending leaflets.

¹Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK. ²Centre for Evolutionary and Functional Ecology, UMR 5175, 1919 Route de Mende, 34293 Montpellier cedex 5, France. ³PO Box 157, Lord Howe Island, New South Wales 2898, Australia. ⁴Department of Systematic Botany, University of Aarhus, DK-8000 Aarhus C, Denmark. ⁵Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland.

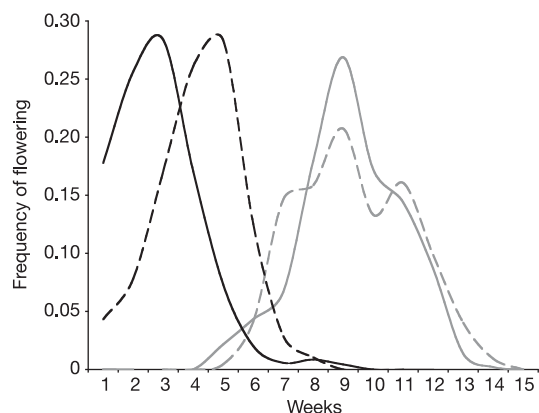


Figure 2 | Flowering phenology for each *Howea* species. *H. belmoreana* is shown in grey ($n = 198$), *H. forsteriana* in black ($n = 177$), with male (solid line) and female (dotted line) phases (see Methods). The flowering times of the two species are strongly displaced. In addition, both species are significantly different in their sexual synchrony index (SI; $F_{1,222} = 27.26$, $P < 10^{-4}$). The SI is not significantly different from zero ($n = 132$, $t = -0.30$ weeks (t -test), $P = 0.18$) and does not differ among sites of *H. belmoreana* ($F_{8,131} = 1.54$, $P = 0.15$), whereas *H. forsteriana* is strongly protandrous (mean SI = 1.3 weeks, $n = 92$, $t = 6.75$, $P < 10^{-4}$) and there are significant differences among sites ($F_{8,83} = 1.96$, $P = 0.0619$).

itself has eroded rapidly and will be awash within 200,000 years. Apart from Ball's Pyramid, a precipitous sea stack 23 km southeast of Lord Howe Island that supports only limited plant life, there are no islands nearby, and Australia is the nearest land mass. However, the indigenous vascular flora of Lord Howe Island has greater affinities with New Zealand and New Caledonia than with Australia. Of its 241 plant species, almost half are endemic, and the terrestrial fauna of the island shows similar levels of endemism. Lord Howe Island represents an ideal location to study sympatric speciation because it has long been isolated, is of known age and is so small that geographical isolation on the island cannot realistically occur³.

The palm family is represented on Lord Howe Island by four species in three strictly endemic genera. The two species of *Howea*, *H. belmoreana* and *H. forsteriana*, are extremely abundant, occurring in more than 70% of the island's vegetation¹⁶. *Howea forsteriana*, the kentia palm, is one of the most widely traded houseplants in the world and is worth over €7 million per year in the Dutch horticultural industry alone. The two species of *Howea* display striking morphological differences and their taxonomic status is indisputable¹⁷ (Fig. 1b, c). They occur sympatrically in numerous places on Lord Howe Island¹⁶, and yet putative hybrids have only rarely been reported¹⁸; our thorough fieldwork identified only five specimens

with intermediate morphologies. We have also confirmed that both species are diploid ($2n = 32$) using conventional cytological techniques, thereby excluding polyploid speciation¹⁹.

We have produced the most comprehensive DNA-based phylogenetic tree for the largest subfamily of palms (Arecoideae), comprising 132 taxa, including all 67 genera of the Indo-Pacific Areceae²⁰. These data strongly support the monophyly of *Howea* and a sister relationship to the monotypic Australian genus *Laccospadix*. Correcting for molecular rate heterogeneity across lineages, we have dated this tree using four calibration points simultaneously, all independent from Lord Howe Island. Using two different molecular dating methods^{21,22}, we estimated the split between *Howea* and *Laccospadix* to be 4.57–5.53 Myr old, and that the two *Howea* species diverged 1.92 ± 0.53 Myr ago (nonparametric rate smoothing (NPRS); Table 1) or just less than 1 Myr ago (bayesian; Table 1), long after Lord Howe Island was formed. Other dates in the tree are also consistent with geological history, including the root nodes of the other two endemic Lord Howe Island palm genera, *Hedyscepe* and *Lepidorrhachis*, and the root node of *Carpoxylon*, which is younger than the age of Vanuatu²³ on which it is endemic (Table 1).

During fieldwork, we monitored both *Howea* species throughout the flowering season. Phenological data indicate that the species are reproductively isolated, with *H. forsteriana* flowering before *H. belmoreana*. The peak flowering of each species is separated by approximately six weeks and has limited overlap (Fig. 2). *Howea forsteriana* is protandrous at the population level, with male flowering peaking two weeks before female receptivity, whereas *H. belmoreana* is synchronous (Fig. 2). Notably, when *H. forsteriana* occurs on volcanic rather than calcareous substrates, asynchronous maturation is not observed (Supplementary Information; $t = 0.49$ (t -test), $n = 12$, $P = 0.63$). Thus, flowering-time differences seem to be directly influenced by substrate-induced physiological changes. We also found that *Howea* is wind-pollinated, a rare syndrome in palms (contrary to popular belief), and complete exclusion experiments demonstrated the absence of apomixis in both species.

The distributions of *H. forsteriana* and *H. belmoreana* are also dependent on soil pH (Fig. 3). *Howea belmoreana* is restricted to neutral and acidic soils, whereas *H. forsteriana* prefers calcarenite, a recent basic sedimentary formation that dominates low-lying parts of the island³. The same pattern is observed for both adults and juveniles ($r^2 = 0.79$, $n = 22$ sites, $P < 10^{-4}$ for *H. forsteriana*; $r^2 = 0.69$, $n = 43$, $P < 10^{-4}$ for *H. belmoreana*). Despite this preference, both species occur sympatrically in 11 out of the 55 quadrats that contain palms (see Methods).

Consistent with sympatric speciation, genetic divergence (F_{ST}) within the genome, estimated using 274 polymorphic amplified fragment-length polymorphism (AFLP) loci, follows an L-shaped distribution, with most loci showing low F_{ST} and only a small number

Table 1 | Ages for root nodes (in Myr) calculated from molecular phylogenetic trees

Taxa	Node number*	Distribution	Node age $\pm$ s.d. (NPRS)	Node age $\pm$ s.d. (bayesian)
<i>Acanthophoenix/Tectiphala</i>	1	Mascarenes†	7.61 $\pm$ 1.70	6.52 $\pm$ 1.10
<i>Dictyosperma</i> (hurricane palm)	2	Mascarenes‡	7.75 $\pm$ 1.83	6.60 $\pm$ 1.22
<i>Hyophorbe</i> (bottle palm)	3	Mascarenes‡	7.75 $\pm$ 0.02	7.45 $\pm$ 0.38
Palms	4	Widespread§	88.93 $\pm$ 9.38	85.73 $\pm$ 2.20
<i>Carpoxylon</i>	5	Vanuatu	7.07 $\pm$ 2.05	11.98 $\pm$ 5.36
<i>Hedyscepe</i> (big mountain palm)	6	LHI	6.66 $\pm$ 2.45	8.24 $\pm$ 4.30
<i>Lepidorrhachis</i> (little mountain palm)	7	LHI	7.77 $\pm$ 3.07	4.62 $\pm$ 3.28
<i>Howea</i>	8	LHI	4.57 $\pm$ 1.45	5.53 $\pm$ 2.89
<i>Howea belmoreana</i> (curly palm)/ <i>Howea forsteriana</i> (kentia palm) split	9	LHI	1.92 $\pm$ 0.53	<1.00¶

LHI, Lord Howe Island.

*Nodes 1–4 are calibration points. See also Supplementary Information.

†*Acanthophoenix* occurs on La Réunion (2 Myr) and Mauritius (7.8 Myr); *Tectiphala* occurs on Mauritius only.

‡*Dictyosperma* and *Hyophorbe* occur on La Réunion, Mauritius and Rodrigues.

§Santonian fossil, *Sabalites magothiensis*, 83.5 Myr old (ref. 30).

|| The age of Vanuatu is about 10 Myr, although the exact figure is disputed²³. Therefore, we did not use this point as a calibration; it is given here for comparative purposes only.

¶ Age under the 1-Myr detection level in the bayesian dating calculation of this data set.

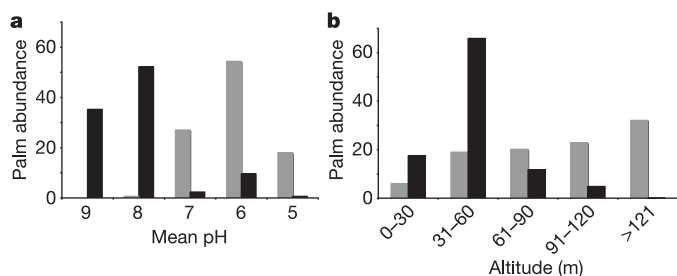


Figure 3 | Distribution of *H. forsteriana* and *H. belmoreana* according to mean soil pH and elevation. Number of palm trees; *H. forsteriana* is shown in black ($n = 1,677$) and *H. belmoreana* in grey ($n = 4,542$). **a, b,** Distribution according to the mean pH of the soil (**a**) in each quadrat and elevation (**b**). *Howea forsteriana* is found in sites at lower elevation ($P < 3 \times 10^{-4}$) and higher soil pH ($P < 10^{-4}$) than *H. belmoreana*, and is largely restricted to calcarenite. However, both species coexist over a large range of elevations and soil pH.

of loci showing high levels of divergence (Fig. 4). Accordingly, the median of F_{ST} is much lower than the mean (median 0.131; mean 0.307 ± 0.020). In the upper tail of the distribution, only four of these AFLP loci differ more strongly between the two species than expected under neutrality^{5,24} (Fig. 4). This signature of species divergence is completely different from that of allopatric speciation, for which genetic differences are expected to accumulate throughout the entire genome²⁵. The four loci were the only markers that were fixed in the two species ($P > 0.95$; Fig. 4), resulting in an experiment-wide significance level for rejection of the null-hypothesis of neutrality of $P < 10^{-5}$: these loci are those most likely to be linked to genes subject to divergent selection during sympatric speciation.

Thus, our phylogenetic, ecological and genomic data are consistent with the following scenario. The ancestor of *Howea* reached Lord Howe Island, most likely from Australia, as long as 4.5–5.5 Myr ago. More recently, *H. forsteriana* diverged from its sister species (an ancestor of *H. belmoreana*) by colonizing widespread lowland calcarenite deposits. Calcarenite dates from the mid-Pleistocene²⁶, which corresponds well to the age of the split between the two *Howea* species recovered from the molecular clocks (Table 1). The absence of protandry in *H. forsteriana* on less basic soils indicates that the conspicuous flowering time difference may have arisen as a physiological response to a new substrate, thus introducing a bias in random mating and kicking-off speciation. However, despite their striking edaphic preferences, it is very unlikely that the two species have ever been truly spatially isolated from each other, given that Lord Howe Island is so small, that volcanic and calcarenite substrates are highly interdigitated, that both species co-occur in ~20% of their modern distribution range and that both are wind-pollinated. The role of other islands on the Lord Howe Rise is not relevant to the speciation event because all of them eroded long before the *Howea* species diverged from each other. The role of the nearest landmasses, Australia, New Caledonia and New Zealand, is equally unlikely to be involved in the speciation of *Howea*, as extensive exploration of the Indo-Pacific palm flora has shown that *Howea* is strictly endemic to Lord Howe Island.

Current models of sympatric speciation invoke either one ‘magic trait’ that is both subject to disruptive selection and controls non-random mating, or linkage disequilibrium between genes controlling assortative mating and those conferring novel adaptation^{2,7,9,27}. We do not know which model is more likely in the case of *Howea*. Changes in flowering time may have arisen as a plastic response to exogenous stress on calcarenite. In this scenario, heritable changes in flowering time may have been induced after the speciation event was complete. Alternatively, speciation may have been facilitated by linkage disequilibrium between genes controlling adaptation to calcarenite and those controlling flowering time. Both traits differ

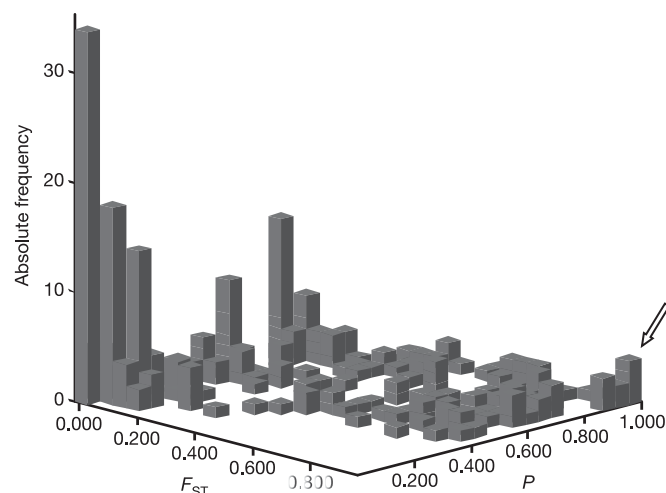


Figure 4 | AFLP genome scan for species differentiation in *H. belmoreana* and *H. forsteriana*. Histogram shows the frequency distribution of estimates for interspecific genetic divergence (F_{ST}) for 274 AFLP loci. Frequencies are on the vertical axis; F_{ST} and the probability of departure from the null-hypothesis of neutrality²⁴ are on the horizontal axes. In the upper tail of the F_{ST} distribution, four loci that depart from neutral expectations at the 0.95 level are indicated by an arrow. These four loci have equal probabilities as all four are fixed in the two species.

between the two palms (Figs 2, 3). In any case, our data are consistent with models of sympatric speciation involving disruptive/divergent selection on a limited number of genes^{2,7,9} (Fig. 4). The data also indicate that the gene pools of these wind-pollinated species are extremely homogeneous, with only 5% of molecular variance between sampled localities for each species (analysis of molecular variance, AMOVA). Hence, mating within the ancestral homogeneous population that gave rise to the two species of *Howea* was indeed random with respect to the place of birth of the mating partners, a characteristic that distinguishes sympatric speciation from other modes of species divergence².

METHODS

Phylogenetic analyses and dating. We have sequenced two low-copy nuclear regions (intron 4 of phosphoribulokinase and intron 23 of RNA polymerase II subunit 2) for Arecoideae, with complete genus level sampling for Areceae²⁰ using polymerase chain reaction (PCR) and DNA automated sequencing (2,464 nucleotides for each of 132 taxa). An initial search using maximum parsimony and 1,000 random taxon-additions and tree bisection-reconnection as implemented in PAUP²⁸ yielded 2,955 most-parsimonious trees. Using ModelTest we identified HKY85+G as a suitable model of DNA evolution and saved one of the most-parsimonious trees with likelihood branch lengths. Following a likelihood ratio test, a constant molecular clock was rejected ($-\ln L c + = 20406.36912$; $-\ln L c - = 20574.10734$; $P < 0.0001$). Therefore, we used both non-parametric rate smoothing (r8s; ref. 21) and bayesian dating methods (DivTime²²) to generate an ultrametric tree, calibrated using four calibration points simultaneously. The root nodes of three independent lineages endemic to the Indian Ocean Mascarene Islands (*Acanthophoenix*/*Tectiphiala*, *Dictyosperma* and *Hyophorbe*) were constrained to be no older than 7.8 Myr (ref. 29). The oldest known palm fossil, *Sabalites magothiensis* from the Santonian, provided a minimum age for the root of the tree of 83.5 Myr (ref. 30). Standard deviations were calculated by re-applying the procedure to 100 bootstrapped DNA matrices²¹.

Habitat ecology. Seventy-eight non-overlapping quadrats (20 × 20 m), were generated at random using geographic information system (GIS). Within each quadrat, we recorded the total number of adults and juveniles of each *Howea* species, and the elevation and soil pH. Juveniles were defined as individuals lacking an aerial stem. Soil samples were collected in three places in each quadrat from a depth of 20 cm below the soil surface. The pH of each sample was measured using Inoculo soil pH test kits (EnviroEquip Pty). Differences in pH and elevation between species were tested with a Student's *t*-test.

Phenology. To determine the phenology of *Howea*, nine sites of each species

distributed across the island were visited weekly (September–December 2003). Within each site, 20–25 reproductive specimens were selected. The inflorescences of *Howea* are monoecious and long-lived, with a delay of one year between male and female anthesis, and more than a year until fruit maturation. Thus, each inflorescence persists for four seasons, reaching male anthesis in season 1, female anthesis in season 2, and with fruits ripening through seasons 3 and 4, and any one individual palm bears functionally male, female and fruiting inflorescences simultaneously. For each palm specimen and at each visit, the number of inflorescences in male, female and fruiting phases was counted (A, pre-anthesis; B, anthesis; C, post-anthesis). A synchrony index (SI) was defined at the plant level as the time between the first week of male phase B and the first week of female phase B. An analysis of variance (ANOVA) was used to test for differences in SI between species and sites. Student's *t*-tests were used to determine whether the SI was significantly different from zero for each species.

Pollination. Flowering spikes nearing female anthesis were subjected to three treatments: (1) insect and wind-borne pollen exclusion (using a paper bag), (2) insect-only exclusion (using a fine-mesh fabric bag), and (3) open pollination. The number of female buds was counted on each spike. Inflorescences were treated well before anthesis, and bags were removed one week after the end of anthesis to prevent a detrimental effect on fruit maturation. Six weeks after the end of female anthesis, the number of developing fruits was counted on each spike. Data were analysed using paired *t*-tests.

Genome scan. AFLP fragments were generated from total DNA using an AFLP plant mapping kit (Applied Biosystems). A total of 48 primer pairs were tested for variability before selecting pairs B13, G7 and Y1 (see manufacturer's protocol). Bands were scored manually using Genotyper 2.0. We estimated F_{ST} values for 274 AFLP bands that matched our polymorphism criteria (Supplementary Information) and compared them to neutral expectations from simulations based on the observed average F_{ST} between the two species. Computer simulations and calculation of significance levels were carried out with the software Dfdist by Beaumont, which uses the method by Beaumont and Nichols²⁴ adapted to dominant markers. Initial simulations indicated that the results remained robust across a wide range of effective population sizes (N_e) and mutation rates (μ), as already reported by in ref. 24, and the final runs were carried out with 50,000 realizations assuming $N_e = 50,000$ and $\mu = 1 \times 10^{-5}$. Outlier loci were identified in a two-step procedure in which highly divergent loci were selected based on a first round of simulations, the average F_{ST} was recalculated without them, and a second round of simulations based on the adjusted average F_{ST} was used to identify outliers subject to directional selection. The shape of the F_{ST} distribution was described by comparing the median with the mean.

Received 23 November 2005; accepted 4 January 2006.

Published online 8 February 2006.

- Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, Sunderland, Massachusetts, 2004).
- Gavrilets, S. Models of speciation: what have we learned in 40 years? *Evolution* **57**, 2197–2215 (2003).
- McDougall, I., Embleton, B. J. J. & Stone, D. B. Origin and evolution of Lord Howe Island, southwest Pacific Ocean. *J. Geol. Soc. Aust.* **28**, 155–176 (1981).
- Luikart, G., England, P., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nature Rev. Genet.* **4**, 981–994 (2003).
- Beaumont, M. Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol. Evol.* **20**, 435–440 (2005).
- Turelli, M., Barton, N. H. & Coyne, J. A. Theory and speciation. *Trends Ecol. Evol.* **16**, 330–343 (2001).
- Dieckmann, U. & Doebeli, M. On the origin of species by sympatric speciation. *Nature* **400**, 354–357 (1999).
- Higashi, M., Takimoto, G. & Yamamura, N. Sympatric speciation by sexual selection. *Nature* **402**, 523–526 (1999).
- Kondrashov, A. S. & Kondrashov, F. A. Interactions among quantitative traits in the course of sympatric speciation. *Nature* **400**, 351–353 (1999).
- Tregenza, T. & Butlin, R. K. Speciation without isolation. *Nature* **400**, 311–312 (1999).
- Filchak, K. E., Roethele, J. B. & Feder, J. L. Natural selection and sympatric divergence in the apple maggot *Rhagoletis pomonella*. *Nature* **407**, 739–742 (2000).
- Schliwien, U. K., Tautz, D. & Pääbo, S. Sympatric speciation suggested by crater lake cichlids. *Nature* **368**, 629–632 (1994).
- Meyer, A., Kocher, T. D., Basasibwaki, K. P. & Wilson, A. C. W. Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature* **347**, 550–553 (1990).
- Bush, G. L. Sympatric speciation in animals, new wine in old bottles. *Trends Ecol. Evol.* **9**, 285–288 (1994).
- Sorenson, M. D., Sefc, K. M. & Payne, R. B. Speciation by host switch in brood indigo birds. *Nature* **424**, 928–931 (2003).
- Pickard, J. Vegetation of Lord Howe Island. *Cunninghamia* **1**, 133–265 (1983).
- Bailey, L. H. *Howea* in cultivation—the sentry palms. *Gentes Herbarum* **4**, 189–198 (1939).
- Oliver, W. R. B. The vegetation and flora of Lord Howe Island. *Trans. Proc. NZ Inst.* **49**, 94–161 (1917).
- Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
- Dransfield, J. *et al.* A new phylogenetic classification of the palm family, *Arecaceae*. *Kew Bull.* (in the press).
- Sanderson, M. J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1231 (1997).
- Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352–361 (2001).
- Burrett, C., Duhig, N., Berry, R. & Varne, R. Asian and south-western Pacific continental terranes derived from Gondwana, and their biogeographic significance. *Aust. Syst. Bot.* **4**, 13–24 (1991).
- Beaumont, M. & Nichols, R. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B* **263**, 1619–1626 (1996).
- Via, S. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol. Evol.* **16**, 381–390 (2001).
- Brooke, B. R., Woodroffe, C. D., Murray-Wallace, C. V., Heijnis, H. & Jones, B. G. Quaternary calcarenite stratigraphy on Lord Howe Island, southern Pacific Ocean and the record of coastal carbonate deposition. *Quat. Sci. Rev.* **22**, 859–880 (2003).
- Doebeli, M., Dieckmann, U., Metz, J. & Tautz, D. What we have also learned: adaptive speciation is theoretically plausible. *Evolution* **59**, 691–695 (2005).
- Swofford, D. L. *PAUP* 4.0: Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Sinauer Associates, Sunderland, Massachusetts, 2001).
- Vaughan, R. E. & Wiehe, P. O. Studies on the vegetation of Mauritius 1: a preliminary survey of the plant communities. *J. Ecol.* **25**, 289–343 (1937).
- Berry, E. W. A palm from the mid-Cretaceous. *Torreya* **5**, 30–33 (1905).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Lord Howe Island Board, L. Wilson, T. Wilson, J. Smith and the New South Wales National Parks and Wildlife Service for facilitating this research; E. Brown and B. Conn for logistical support; T. Wilmot, R. Cowan, L. Llédó, I. Gallego and L. Hanson for research assistance; C. Lewis, J. Roncal, M. Thomas and M. Beaumont for sharing data and software; J. Dowe, the Flecker Botanic Garden and Royal Botanic Gardens, Sydney for providing material; J. Moat for GIS assistance; J. Dransfield, J.-C. Pintaud, S. Gavrilets, M. Chase, T. Barraclough and L. Rieseberg for discussions; and the Royal Botanic Gardens, Kew, the BAT Biodiversity Partnership, the European Commission and the Leverhulme Trust for funding.

Author Information DNA sequences have been deposited at EBI under accession numbers AF453329–AF453381, AY348907–AY348944, AY543096–AY5443156, AJ830020–AJ831373, AJ971821–AJ971841 (see Supplementary Information). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.S. (v.savolainen@kew.org).

LETTERS

Future fitness and helping in social queues

Jeremy Field¹, Adam Cronin¹ & Catherine Bridge¹

Helpers in primitively eusocial and cooperatively breeding animal societies forfeit their own reproduction to rear the offspring of a queen or breeding pair, but may eventually attain breeding status themselves. Kin selection¹ provides a widely accepted theoretical framework for understanding these societies, but differences in genetic relatedness do not explain a universal societal feature: the huge variation between individuals in helping effort^{2–10}. An alternative explanation for this variation lies in a fundamental trade-off faced by helpers: by working harder, they increase the indirect component of their fitness, but simultaneously decrease their own future survival and fecundity^{2,4,8}. Here, we show that individuals work less hard when they stand to lose more future fitness through working. We experimentally manipulated two components of future fitness in social queues of hover wasps (Stenogastrinae): a helper's chance of inheriting an egg-laying position, and the workforce available to rear her offspring should she inherit. After each manipulation, helpers increased or decreased their effort as appropriate to the change in expected future fitness that they experienced. Although helping provides significant indirect fitness benefits for hover wasps¹¹, our study shows that variation in the costs associated with helping is the major determinant of helping effort.

A conspicuous general feature of primitively eusocial and cooperatively breeding societies is the enormous variation between individuals in helping effort^{2–4}. Initial attempts to understand this variation focused on the prediction that more help should be given to closer relatives⁵. Yet tests of this hypothesis have produced mixed results. In eusocial insects, there is little evidence of nepotism within colonies, except when obvious cues like sex of the offspring are correlated with differences in relatedness^{6,7,12}. Kin discrimination is widespread in cooperatively breeding vertebrates, but on average only 10% of the variation in helping effort can be explained by variation in relatedness⁵. This lack of explanatory power has led some to question whether kin selection provides a general explanation for helping^{9,10,13}. In this study, we test experimentally whether variation in helping effort is driven primarily by variation in the costs of helping^{2–4,8,14}.

By working harder to rear the offspring of a relative, helpers simultaneously decrease their own future survival and the fecundity they can expect through inheriting breeding positions themselves^{2–4,8,15}. This life-history framework suggests that individuals should balance current levels of help against anticipated future fitness returns. Social queues, where group members inherit breeding positions in a predictable, temporally stable order, are widespread in both insects and vertebrates^{16–18}. Queues lead to systematic individual variation in future fitness. First, individuals nearer to the front of the queue have a shorter wait to inherit, and therefore a greater chance of surviving to breed^{14,8,16,17}. Second, if breeders are more productive in larger groups, individuals waiting in longer queues can expect greater reproductive success should they succeed in inheriting breeding positions^{4,8,16}. If non-breeders adjust their helping effort according to their expected future fitness, we can make

two predictions. First, high-ranking individuals should work less hard than low-rankers. Second, individuals at a given rank in the queue should work less hard in larger groups. In both cases, individuals work less hard when they expect to lose more future fitness through working. We test these predictions experimentally in the tropical hairy-faced hover wasp *Liostenogaster flavolineata* Cameron, a system in which both rank and group size can be manipulated.

L. flavolineata (Hymenoptera: Stenogastrinae) nests in groups of 1–10 related¹⁹ females (within-nest coefficient of relatedness in this study $r = 0.46 \pm 0.08$). A single female, known as the dominant or rank 1, lays almost all of the eggs and rarely leaves the nest¹⁹. She is morphologically undifferentiated²⁰ from other group members, known as helpers (rank 2 and upwards). Helpers provide aid by foraging to collect food for the dominant's immature larvae, and helping effort is conveniently measured by time spent away from the nest⁴. Number of offspring reared is positively correlated with group size^{11,16}, so that helping provides indirect fitness benefits. Dominance in *L. flavolineata* is determined through strict age-based queuing, with rank 1 being the oldest female. When rank 1 dies or is experimentally removed, the next-oldest female becomes dominant in 90% of cases²¹. By knowing their relative ages, it is thus possible to order females precisely according to their positions in the queue. Reduced helping effort by high-ranking females could be adaptive, because females reaching rank 1 experience a significant increase in expected lifespan^{16,18} and rear offspring to which they are more closely related¹¹.

Consistent with our predictions based on future fitness, in unmanipulated *L. flavolineata* groups, lower-ranked helpers and helpers in smaller groups worked harder (Fig. 1 and see Methods). As in other systems⁹, a helper's genetic relatedness to the dominant egg layer had no effect on her helping effort, despite considerable variation in relatedness (Fig. 1 inset). In order to test whether future fitness determines helping effort causally, we experimentally manipulated two components of the direct reproduction that a helper could expect in the future. In experiment 1, we tested whether chance of inheritance determines helping effort by experimentally promoting females that lay third in the queue (Fig. 2a). We removed the second-oldest (rank 2) female from each experimental group, so that focal rank 3 helpers were promoted to rank 2. We controlled for the resulting change in group size by removing a low-ranking female (below rank 3) from a second set of matched groups. As predicted, experimental rank 3 helpers, whose chance of inheriting had been increased by the manipulation, afterwards worked significantly less hard than control rank 3 helpers, whose chance of inheriting remained unchanged (Fig. 3a). Before being promoted, experimental rank 3 helpers had worked significantly harder than control rank 2 helpers ($62 \pm 13\%$, $n = 10$ versus $11 \pm 3\%$, $n = 13$ time off the nest, $P = 0.002$). After being promoted to rank 2 themselves, however, experimental rank 3 helpers worked approximately as hard as control rank 2 helpers ($27 \pm 8\%$ versus $20 \pm 6\%$, $P = 0.5$). Our manipulation altered the relationship between helping effort and age, in

¹Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.

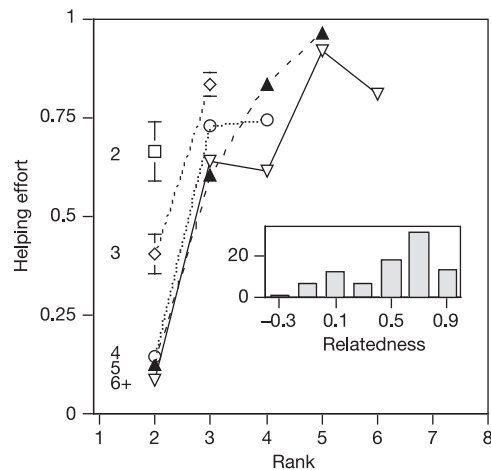


Figure 1 | Helping effort as a function of rank and group size in unmanipulated *L. flavolineata* groups. Each line represents a different group size, as indicated by the numbers on the left of the graph (open square, 2; open diamond, 3; open circle, 4; filled triangle, 5; open triangle, 6+). Helping effort is estimated as the proportion of censuses spent away from the nest. For clarity, $\pm$ s.e.m. is shown for only the two smallest group sizes, which include the largest and smallest errors. In an analysis of covariance using the full data set of 332 helpers (97 nests), helper rank, group size and helper absolute age had significant ($P < 0.0001$) effects on effort, together accounting for 45% of the total variation. The effect of absolute age was that younger helpers worked harder. Helper-dominant relatedness had no effect ($P > 0.5$) when the analysis was repeated using a subset of 88 helpers from 2001 for which genotypes were available. Inset gives the frequency distribution of the 88 helper-dominant raw relatedness values, showing peaks close to the expected values for cousins ($r = 0.1875$) and sisters ($r = 0.75$). See Methods for further details.

which younger (lower-ranked) helpers normally work harder than older (higher-ranked) helpers. Despite being significantly younger than control rank 2 helpers (55 ± 13 days old versus 116 ± 10 days old, $P = 0.001$), promoted rank 3 helpers still behaved as expected for their new rank, suggesting that rank itself, rather than age or any other correlate, has a causal effect.

In experiment 2, we tested the response of rank 2 helpers to a second component of future fitness: the workforce available to rear their offspring should they succeed in inheriting. We experimentally reduced the workforce by removing all but ranks 1 and 2 from groups where there were initially 3–5 females (Fig. 2b). In order to avoid leaving reduced groups with more dependent offspring per helper than unmanipulated controls¹¹, we also removed a proportion, R/N , of the immature larvae present on experimental nests, where R is the number of helpers removed and N is the initial number of helpers. Although the ratio of helpers/dependent offspring did not differ between treatments after the manipulation, rank 2 helpers on nests where group size had been reduced worked harder than control rank 2 helpers (Fig. 3b). This agrees with our prediction based on future fitness: both sets of rank 2 helpers retained the same chance of inheriting breeding positions, but rank 2 helpers in reduced groups could expect less direct fitness after inheriting because they would have fewer helpers. The change in behaviour of experimental rank 2 helpers is unlikely to represent an abnormal response to disturbance caused by the manipulation. We allowed a 2-day gap between the manipulation and measurement of post-manipulation helping effort. Also, after their groups had been reduced, experimental rank 2 helpers spent a similar proportion of time off the nest ($54 \pm 10\%$, $n = 8$) as rank 2 helpers in unmanipulated groups of two females ($47 \pm 9\%$, $n = 7$).

Our results suggest that the trade-off between current levels of help and future direct fitness through inheritance is a major determinant of helping effort in *L. flavolineata*. Helpers can respond to

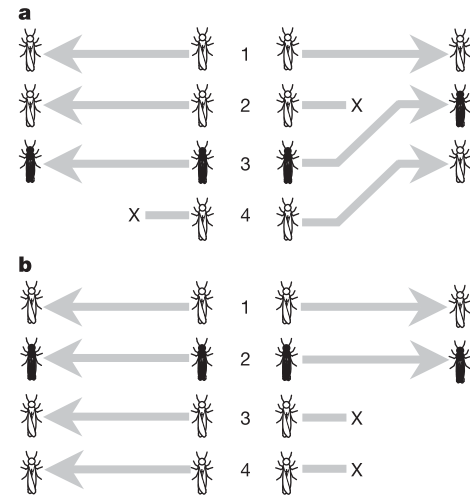


Figure 2 | Experimental design with a group size of 4 females. Numbers indicate ranks (rank 1 is the dominant). The right-hand side of each figure represents nests where the rank (experiment 1; **a**) or group size (experiment 2; **b**) of the focal wasp (in black) was manipulated, whereas the left-hand side represents control nests. On each side, the inner column of 4 wasps shows the pre-manipulation state and the outer column the post-manipulation state, where crosses signify individuals that were removed in the manipulations. Note that experiment 1 can be used to draw conclusions about the effect of rank but not group size, as group size was manipulated to the same extent in both treatments.

changes in their position in the queue or the size of their group by either increasing or decreasing their effort as appropriate. Our results imply that two other processes that could modify helping effort—group augmentation and compensation—are less important in *L. flavolineata*. Group augmentation occurs when an individual can boost group size by helping, so that she will later have more helpers herself if she survives to inherit²². However, group augmentation effects should have led to the rank-promoted females in experiment 1 working harder rather than less hard, because the manipulation increased their chance of inheriting and receiving help from any offspring that they contributed to rearing. Our result suggests that such benefits are outweighed by the negative effect that working harder would have on the chance of inheritance itself.

L. flavolineata helpers respond differently to changes in how hard other group members work compared with birds, where the removal of helpers causes remaining adults to compensate by increasing their own helping effort²³. In experiment 1, compensation would have required both control and experimental rank 3 helpers to increase their effort in response to the larger number of dependent offspring per helper after the manipulation. Instead, the control groups showed no compensation, and the experimental groups decreased their effort (Fig. 3a). A key difference may be that clutch size cannot easily be adjusted in many birds, so that the death of a helper increases the benefits of additional help for remaining individuals. In contrast, wasps can respond to helper removals by recycling excess offspring: older excess offspring left after the death of a helper are probably reared by feeding smaller offspring to them, so that investment can be preserved without increasing foraging^{11,24}.

Variation in effort among *L. flavolineata* helpers is a response to variation in the cost parameter (c) in Hamilton's rule¹ ($rb > c$). Although helpers clearly gain indirect fitness benefits through aiding natal nest-mates, informational constraints⁶ may prevent them from responding to variation in relatedness within the group. In contrast, rank and group size may provide easily discernible indications of the future fitness that a helper stands to lose, allowing levels of effort to be behaviourally fine-tuned. Rank and group size should influence an individual's propensity to perform any action that jeopardizes its future fitness, and should be part of any

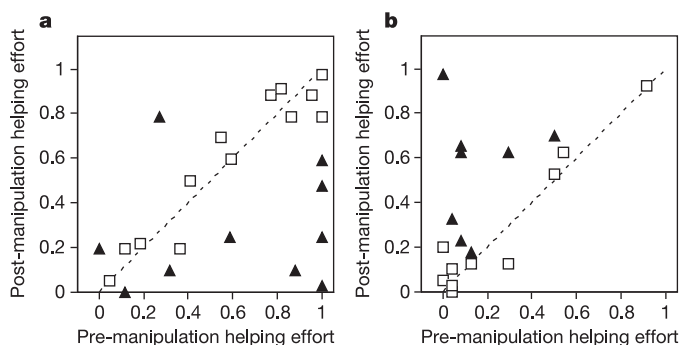


Figure 3 | Results from manipulating the expected future fitness of *L. flavolineata* helpers. In each graph, filled triangles are focal individuals whose rank (experiment 1; **a**) or group size (experiment 2; **b**) was manipulated, whereas open squares are controls. Each data point represents the focal wasp from a different nest. The dashed lines have slope = 1, indicating no change in effort as a result of the manipulations. **a**, In experiment 1, both treatment ($P = 0.01$) and pre-manipulation helping effort ($P = 0.01$) had significant effects on post-manipulation helping effort ($n = 23$). There was also a significant interaction between treatment and pre-manipulation helping effort ($P = 0.003$): focal rank 3 helpers that were already making little effort before the manipulation could not decrease their effort much further after the manipulation. **b**, In experiment 2, both treatment ($P = 0.003$) and pre-manipulation helping effort ($P = 0.002$) had significant effects on post-manipulation helping effort ($n = 18$).

framework for understanding patterns of individual behaviour in social hierarchies⁸.

METHODS

Each experiment used nests situated in culverts (sites) that carried streams under a 4-km stretch of road surrounded by forest between Raub and Bukit Fraser in Pahang State, peninsular Malaysia^{11,16}. Experiment 1 was conducted in late August 2001 and experiment 2 in late August 2003. The relative ages of wasps within each group were known as a result of intensive monitoring for the 5 months before each experiment, during which each newly emerged female was individually marked soon after she reached adulthood¹⁸. In groups where the relative age of the two oldest females was unknown because they had both been present from the start of monitoring, the rank 1 could be identified behaviourally^{18,20}. We allocated nests randomly to treatments after blocking for site, total number of immature offspring and group size (4–8 females in experiment 1; 3–5 females in experiment 2).

Each experiment consisted of three phases. First, all nests were censused every 30 min during the main foraging periods (7.00–11.00) on three consecutive days (days 1–3), for a total of 22 (experiment 1) or 24 (experiment 2) censuses. Each individual's pre-manipulation helping effort was estimated as the proportion of censuses on which she was away from the nest. The second phase was the manipulation itself, carried out on day 4 (experiment 1) or day 6 (experiment 2). All residents on all nests were captured before dawn; subsequently all were released except for the helpers to be removed. In experiment 2, we removed immature offspring from experimental nests at the same time as removing adults. Offspring were divided into three categories: eggs/tiny larvae, medium larvae and large larvae. A proportion, R/N , of each category was removed using a pooter. Pupae, which do not require feeding, were not removed. The final phase of each experiment was estimation of post-manipulation helping effort, on consecutive days 7–10 (experiment 1) and 8–12 (experiment 2), using half-hourly censuses as before. A few groups had to be omitted from the final analysis because group size changed naturally during the course of the experiment or the focal female died or moved up in rank naturally.

Statistical analyses. Data were analysed using generalized linear modelling in the R statistical package (version 2.0.0 for Macintosh)²⁵. The data in Fig. 3 were analysed with y = the focal female's post-manipulation helping effort (that is, her proportion of censuses off the nest). Explanatory variables tested were treatment and the focal female's pre-manipulation helping effort. In the analysis of Fig. 1, explanatory variables tested were helper rank, group size, site, number of immature offspring, helper wing length, helper absolute age and year (2001 or 2003). Rank was correlated with group size ($r^2 = 0.32$; individuals at position n in the queue can occur only in groups of at least size n) and with absolute age ($r^2 = 0.33$; higher-ranked individuals tend to be older). We included all helpers

for which data were available for all explanatory variables ($n = 169$ helpers from 47 nests in 2001; $n = 163$ helpers from 50 nests in 2003; these included the nests that were subsequently manipulated). The same variables had significant effects when data from each year were analysed separately. The effect of relatedness was tested using a subset of 88 of the helpers from 2001. Relatedness was estimated from genotypes at three hyper-variable microsatellite loci (loci and methods as described previously¹⁹). Relatedness had no significant effect whether it was coded as a raw value; a binary variable where 1 = helper significantly ($P < 0.05$) more likely to be a sister than a cousin of the dominant ($n = 48$ individuals) and 0 = this was not the case ($n = 40$); or a binary variable where 1 = helper significantly ($P < 0.05$) more likely to be a sister than a niece of the dominant ($n = 35$) and 0 = this was not the case ($n = 53$). Raw values were estimated using the computer program Relatedness version 5.08 (<http://www.gsofnet.us/GSoft.html>)²⁶. Tests of sister–cousin and sister–niece relationships were conducted using the computer program Kinship version 1.3.1 (<http://www.gsofnet.us/GSoft.html>)²⁷ with a power of 93% (sister–cousin) or 76% (sister–niece) at the $\alpha = 0.05$ level.

We began all analyses with all explanatory variables fitted. Starting with interaction terms, we then subtracted terms from the model until further removals led to significant ($P < 0.05$) increases in deviance, as assessed from tabulated values of F with normal errors or χ^2 with binomial errors²⁸. We report significance levels for terms when adding them last to this minimal adequate model. When there was significant overdispersion using binomial errors, we re-fitted the model assuming a quasibinomial error structure^{25,28}. In each statistical analysis where the dependent variable was helping effort (proportion of censuses spent away from the nest), the results were quantitatively almost identical whether the error structure assumed was binomial or the dependent variable was arcsine-square root transformed then normal errors assumed. Means $\pm$ standard errors are reported.

Received 26 October 2005; accepted 5 January 2006.

- Hamilton, W. D. The genetical evolution of social behaviour. I and II. *J. Theor. Biol.* **7**, 1–52 (1964).
- Heinsohn, R. & Legge, S. The cost of helping. *Trends Ecol. Evol.* **14**, 53–57 (1999).
- Clutton-Brock, T. H. *et al.* Individual contributions to babysitting in a cooperative mongoose, *Suricata suricatta*. *Proc. R. Soc. Lond. B* **267**, 301–305 (2000).
- Cant, M. A. & Field, J. Helping effort and future fitness in cooperative animal societies. *Proc. R. Soc. Lond. B* **268**, 1959–1964 (2001).
- Griffin, A. S. & West, S. A. Kin discrimination and the benefits of helping in cooperatively breeding vertebrates. *Science* **302**, 634–636 (2003).
- Keller, L. Indiscriminate altruism: unduly nice parents and siblings. *Trends Ecol. Evol.* **12**, 99–103 (1997).
- Queller, D. C. *et al.* Unrelated helpers in a social insect. *Nature* **405**, 784–787 (2000).
- Cant, M. A. & Field, J. Helping effort in a dominance hierarchy. *Behav. Ecol.* **16**, 708–715 (2005).
- Clutton-Brock, T. H. Breeding together: kin selection and mutualism in cooperative vertebrates. *Science* **296**, 69–72 (2002).
- Griffin, A. S. & West, S. A. Kin selection: fact and fiction. *Trends Ecol. Evol.* **17**, 15–21 (2002).
- Field, J., Shreeves, G., Sumner, S. & Casiraghi, M. Insurance-based advantage to helpers in a tropical hover wasp. *Nature* **404**, 869–871 (2000).
- Sundstrom, L., Chapuisat, M. & Keller, L. Conditional manipulation of sex ratio by ant workers: a test of kin selection theory. *Science* **274**, 993–995 (1996).
- Stiver, K. A., Dierkes, P., Taborsky, M., Gibbs, H. L. & Balshine, S. Relatedness and helping in fish: examining the theoretical predictions. *Proc. R. Soc. Lond. B* **272**, 1593–1599 (2005).
- Heinsohn, R. & Cockburn, A. Helping is costly to young birds in cooperatively breeding white-winged choughs. *Proc. R. Soc. Lond. B* **256**, 293–298 (1994).
- West-Eberhard, M. J. in *Natural Selection and Social Behaviour: Recent Research and New Theory* (eds Alexander, R. D. & Tinkle, D. W.) 3–17 (Chiron Press, New York, 1981).
- Shreeves, G. & Field, J. Group size and direct fitness in social queues. *Am. Nat.* **159**, 81–95 (2002).
- Kokko, H. & Johnstone, R. A. Social queuing in animal societies: a dynamic model of reproductive skew. *Proc. R. Soc. Lond. B* **266**, 571–578 (1999).
- Field, J., Shreeves, G. & Sumner, S. Group size, queuing and helping decisions in facultatively eusocial hover wasps. *Behav. Ecol. Sociobiol.* **45**, 378–385 (1999).
- Sumner, S., Casiraghi, M., Foster, W. & Field, J. High reproductive skew in tropical hover wasps. *Proc. R. Soc. Lond. B* **269**, 179–186 (2002).
- Field, J. & Foster, W. Helping behaviour in facultatively eusocial hover wasps: an experimental test of the subfertility hypothesis. *Anim. Behav.* **57**, 633–636 (1999).
- Bridge, C. A. L. *Rank and Inheritance in a Facultatively Eusocial Hover Wasp*. Ph.D thesis, Univ. Col., London (2005).
- Kokko, H., Johnstone, R. A. & Clutton-Brock, T. H. The evolution of cooperative breeding through group augmentation. *Proc. R. Soc. Lond. B* **268**, 187–196 (2001).

23. Houston, A. I., Székely, T. & McNamara, J. M. Conflicts between parents over care. *Trends Ecol. Evol.* **20**, 33–38 (2005).
24. Shreeves, G., Cant, M. A., Bolton, A. & Field, J. Insurance-based advantages for subordinate co-foundresses in a temperate paper wasp. *Proc. R. Soc. Lond. B* **270**, 1617–1622 (2003).
25. The R project for statistical computing. (<http://www.R-project.org/>) Vienna Univ. Technol., (2005).
26. Queller, D. C. & Goodnight, K. F. Estimating relatedness using genetic markers. *Evol. Int. J. Org. Evol.* **43**, 258–275 (1989).
27. Goodnight, K. F. & Queller, D. C. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Mol. Ecol.* **8**, 1231–1234 (1999).
28. Crawley, M. J. *Statistical Computing: An Introduction to Data Analysis Using S-Plus* (John Wiley, Chichester, 2002).

Acknowledgements We thank S. Brace, S. Rantala, G. Shreeves and J. Vulliamy for help with fieldwork; K. Durai, H. Rosli and A. Sofian for their hospitality in Malaysia; M. Cant for comments and the Natural Environment Research Council for funding.

Author Contributions C.B. and J.F. did the fieldwork for experiment 1; A.C. and J.F. did the fieldwork for experiment 2; C.B. obtained the genotypic data. J.F. carried out the statistical analyses and wrote the paper. All authors discussed the results and commented on the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.F. (jeremy.field@ucl.ac.uk).

LETTERS

Specification of the neural crest occurs during gastrulation and requires Pax7

Martín L. Basch^{1†}, Marianne Bronner-Fraser¹ & Martín I. García-Castro^{1,2}

The neural crest is a stem population critical for development of the vertebrate craniofacial skeleton and peripheral ganglia. Neural crest cells originate along the border between the neural plate and epidermis¹, migrate extensively and generate numerous derivatives, including neurons and glia of the peripheral nervous system, melanocytes, bone and cartilage of the head skeleton². Impaired neural crest development is associated with human defects, including cleft palate³. Classically, the neural crest has been thought to form by interactions at the border between neural and non-neural ectoderm or mesoderm^{4–7}, and defined factors such as bone morphogenetic proteins (BMPs) and Wnt proteins have been postulated as neural crest-inducers^{8–13}. Although competence to induce crest cells declines after stage 10 (ref. 14), little is known about when neural crest induction begins *in vivo*. Here we report that neural crest induction is underway during gastrulation and well before proper neural plate appearance. We show that a restricted region of chick epiblast (stage 3–4) is specified to generate neural crest cells when explanted under non-inducing conditions. This region expresses the transcription factor Pax7 by stage 4+ and later contributes to neural folds and migrating neural crest. In chicken embryos, Pax7 is required for neural crest formation *in vivo*, because blocking its translation inhibits expression of the neural crest markers Slug, Sox9, Sox10 and HNK-1. Our results indicate that neural crest specification initiates earlier than previously assumed, independently of mesodermal and neural tissues, and that Pax7 has a crucial function during neural crest development.

To investigate early events in chick neural crest formation, we analysed the expression of known neural crest markers at progressively earlier embryonic stages. Expression of the paired box transcription factor Pax7 (ref. 15) (Fig. 1a) correlates with the presumptive neural crest domain in gastrulating embryos and precedes known neural crest markers such as Slug¹⁶ and SoxE¹⁷. Using *in situ* hybridization and immunostaining, the Pax7 expression domain first appears at stage 4+ as two bilaterally symmetric oblique bands lateral to Hensen's node, extending diagonally about 400 μ m towards the primitive streak. At stage 4+, Pax7 expression does not overlap with markers of presumptive epidermis (*Gata2*), neural plate (*Sox2*), ectodermal placode (*Dlx5*) or neural plate border (*BMP4* and *Msx1*)^{18–20}. As the neural plate forms, Pax7 expression partly overlaps the border (stage 5; Fig. 1c), and from stage 6 onwards gradually shifts to superimpose on the neural folds (with the exception of forebrain that forms no neural crest) (Fig. 1a). In contrast, the related Pax3 gene was expressed more caudally and medially²¹ (Fig. 1b). Later, Pax7 is expressed by migrating neural crest cells at midbrain and hindbrain regions (Fig. 2a). Focal injections of CM-DiI were used to test the fate of the presumptive Pax7 expression domain in stage 3–4 embryos. This region becomes incorporated into the dorsal neural folds (45/57) and, at stages of neural crest

cell migration (17/17), into migrating neural crest cells^{22,23} (Fig. 2b).

We next examined whether the prospective Pax7 region is specified to generate neural crest cells, where 'specification' is defined as the ability to form neural crest in the absence of exogenous signals when placed in a neutral environment. Accordingly, a thin strip of epiblast tissue was dissected perpendicular to, and 250 μ m caudal to, the most rostral extent of the primitive streak of stage 3–4 embryos. The lower layer (hypoblast/mesendoderm) was removed and the epiblast strip was cut into approximately 80- μ m pieces and explanted onto collagen gels. Explants were numbered from 1 to 6, from most lateral to the primitive streak. Only medial explants (about 300 μ m from the primitive streak) generated migratory neural crest cells (staining positively for HNK-1 and Pax7; HNK-1⁺Pax7⁺) ($n = 60/69$ stage 4 embryos (Fig. 2c); $n = 25/30$ stage 3 embryos (Fig. 2d)).

To determine whether these explants could generate proper neural crest derivatives, stage 3 explants incubated for 36 h as above were further cultured under conditions permissive for melanogenesis. After 8 days, medial explants but not adjacent epiblast formed melanocytes (8/9), a genuine neural crest derivative (Fig. 2e), as well as neurons. These results indicate that the presumptive Pax7 domain is specified to form neural crest as early as stage 3, before overt Pax7 expression, without signals from surrounding tissues.

Present models of neural crest induction indicate that interactions between neural and non-neural ectoderm and/or mesoderm might be crucial to the generation of neural crest. If true, this predicts that neural, non-neural and/or mesodermal tissues should be present in medial explants that will express Pax7 and generate neural crest cells. However, definitive neural tissue does not appear until after stage 4+, and it requires neural stabilizing signals emanating from the node and head mesoderm. We therefore examined whether node or mesodermal markers were present within Pax7-expressing medial explants. First we investigated Pax7 protein and mRNA expression on epiblast explants cultured for different durations. Strong Pax7 expression was found on medial epiblast explants as early as 10 h in culture (6/6; Fig. 2f). To examine whether Pax7 expression depended on the presence of mesoderm, Brachyury and Tbx6L (important markers of the node and early mesoderm) were examined together with Pax7 at 10 h. Pax7 protein was detected in medial epiblast explants independent of Brachyury, whereas primitive streak-derived explants were Brachyury-positive (12/12; Fig. 2g). Similarly, most medial epiblast explants expressed Pax7 protein (Fig. 2h) or Pax7 mRNA (Fig. 2i) but failed to express Tbx6L mRNA (5/6 and 3/4 explants, respectively).

Expression of Pax7 independently of mesodermal and neural markers was confirmed on individual stage 3 epiblast explants cultured for 10 h using polymerase chain reaction with reverse transcription (RT-PCR) (Supplementary Fig. 1). Explant sets derived from six embryos were analysed for multiple products in single reactions. Pax7, but not the mesodermal markers Brachyury or

¹Division of Biology 139-74, California Institute of Technology, Pasadena, California 91125, USA. ²MCDDB Yale University, PO Box 208103, New Haven, Connecticut 06520-8103, USA. [†]Present address: House Ear Institute, CMB Department, 2100 W 3rd Street, Los Angeles, California 90029, USA.

Slug (expressed in primitive streak at these early stages), was detected in medial explants. Instead, mesodermal markers were clearly detected in explants derived from primitive streak. *Sox2* (a neural marker) was not detected in stage 3 epiblast explants cultured for 10 h even when derived from anterior regions fated to generate neural plate, but robust expression occurred in neural plate at stage 5 and whole

embryos at stage 10. These findings confirm our *in situ* hybridization and immunostaining results and support a model for early neural crest formation—including *Pax7* expression—independently of mesodermal and neural tissues.

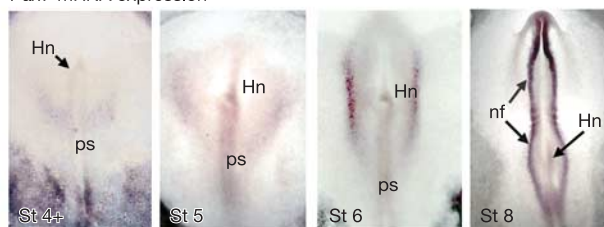
As previous studies have implicated BMP and Wnt signals in neural crest induction, we implanted beads coated with Wnt or BMP inhibitors into the presumptive *Pax7* expression domain. Embryos receiving beads coated with a Wnt inhibitor (Fz5) or with a BMP inhibitor (Noggin) showed a decrease in *Pax7* expression (14/21 with Fz5; 8/10 with noggin), whereas control beads coated in bovine serum albumin (BSA) had no effect (13/13; Supplementary Table 1). These results indicate that Wnt and BMP proteins might be required for neural crest induction during gastrulation.

To examine whether *Pax7* is required for neural crest formation *in vivo*, we used antisense morpholino oligonucleotides (*Pax7*-mo) specifically to prevent the translation of *Pax7* protein. We first showed that *Pax7*-mo reduced translation of *Pax7* protein *in vitro* (Fig. 3a). We then electroporated *Pax7*-mo unilaterally into the presumptive neural crest domain of stage 4 embryos *in vivo*; the contralateral side served as an internal control. Cells that inherited *Pax7*-mo showed a marked decrease in *Pax7* protein levels (Fig. 3b). Furthermore, significant decreases in the neural crest markers *Slug* (24/37 embryos), *Sox9* (40/52) and *Sox10* (36/55) were observed on the electroporated side of the embryos (Fig. 3f–h). Downregulation of *Slug* and *Sox9* in *Pax7*-mo-treated embryos was detected as early as stage 8, corresponding to 8–12 h after electroporation and when *Slug* and *Sox9* are first detectable in the neural folds. In contrast, embryos electroporated with a five-mismatch control morpholino showed no obvious alterations in levels of *Slug* (1/31 embryos), *Sox9* (1/45) or *Sox10* (0/38) (Fig. 3c–e). Cell viability as assessed by staining with 4',6-diamidino-2-phenylindole was unchanged in control and *Pax7*-mo-treated embryos (data not shown). To quantify the effects on neural crest markers, the intensity of expression was scored in identical boxed areas on control and treated sides of embryos that received *Pax7*-mo or control morpholinos. *Sox9* expression was decreased by *Pax7*-mo to less than 40% of that on the contralateral side, but was unchanged in those receiving control morpholino (Fig. 3i). In *Pax7*-mo-treated embryos allowed to develop to stages of cranial neural crest migration, there was a strong decrease in the number of *Pax7*⁺HNK1⁺ migratory midbrain neural crest cells (4/4; Fig. 3j). These results support an important role for *Pax7* in neural crest formation during gastrulation or early neurulation that subsequently affects the migrating population. In contrast to our results for *Pax7*, we were unable to detect an early requirement for *Pax3* using a morpholino against *Pax3* (0/10 embryos showing changes in *Slug* expression, 1/9 for *Sox9* and 2/9 for *Sox10*).

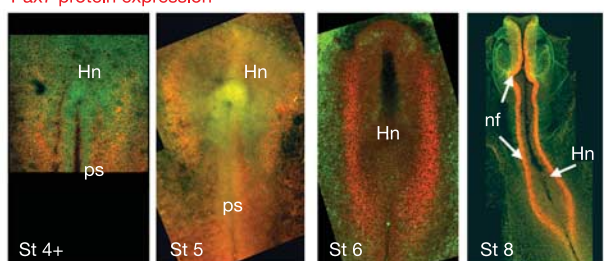
In the mouse, mutations in *Pax3* (ref. 24) and *Pax7* (ref. 25) affect different neural crest derivatives, with *Pax7*^{-/-} mutants showing more rostral alterations than *Pax3*^{-/-} mutants. Extensive overlap in paralogue expression and function was thought to render milder phenotypes²⁶ in single mutants, and double mutants have not been analysed for neural crest phenotypes²⁷. Mouse *Pax3* and *Pax7* expression have been reported to initiate after formation of the neural plate and neural folds at embryonic day (E)8.5 (refs 24, 25). We therefore examined their expression (Supplementary Fig. 2), and found low levels of *Pax3* and *Pax7* in distinct patterns in the epiblast at E6.5. Specifically, we found nuclear *Pax7* in the embryonic ectoderm (future neuroectoderm) and in the neuroepithelium of the future headfold. Nuclear *Pax3* and *Pax7* were detected in anterior neural fold cells by E7.5. Caudal neural folds containing the most recently formed prospective neural crest express only *Pax7*. These data raise the possibility that the mechanisms we describe for birds might be common to amniotes and perhaps all vertebrates. Consistent with this possibility is the observation that knockdown of *Pax3* in *Xenopus* alters neural crest induction²⁸.

Neural crest cells are derived from ectoderm, and were thought to originate by means of interactions between neural and non-neural

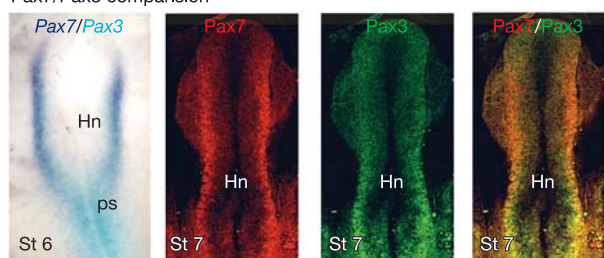
a *Pax7* mRNA expression



Pax7 protein expression



b *Pax7*/*Pax3* comparison



c *Pax7* and epi-neural and border markers

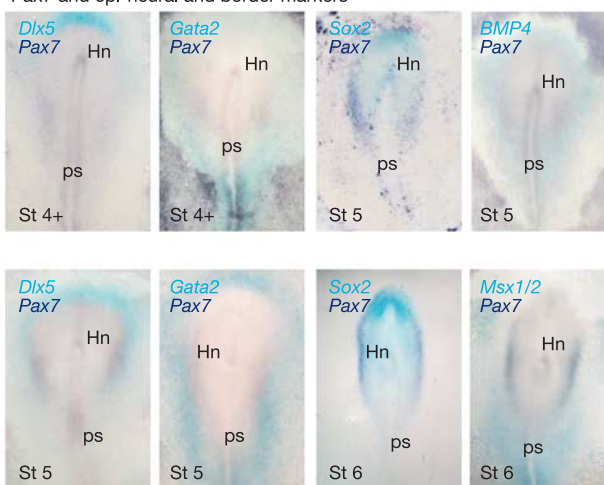


Figure 1 | Early expression of *Pax7* in presumptive neural crest from gastrulation to neurulation. **a**, *In situ* hybridization (top) and immunostaining (bottom, red) for *Pax7* shown at stages (St) 4+, 5, 6 and 8. **b**, Left panel shows double *in situ* hybridization for *Pax7* (purple) and *Pax3* (cyan) in a stage 6 embryo. Right panels, *Pax7* (red), *Pax3* (green) and *Pax7*/*Pax3* immunostaining at stage 7. **c**, Double *in situ* hybridizations comparing *Pax7* expression (purple) with the expression of early neural, non-neural and border markers *Dlx5*, *Gata2*, *Sox2*, *BMP4* and *Msx1* and *Msx2* (cyan). Hn, Hensen's node; nf, neural folds; ps, primitive streak.

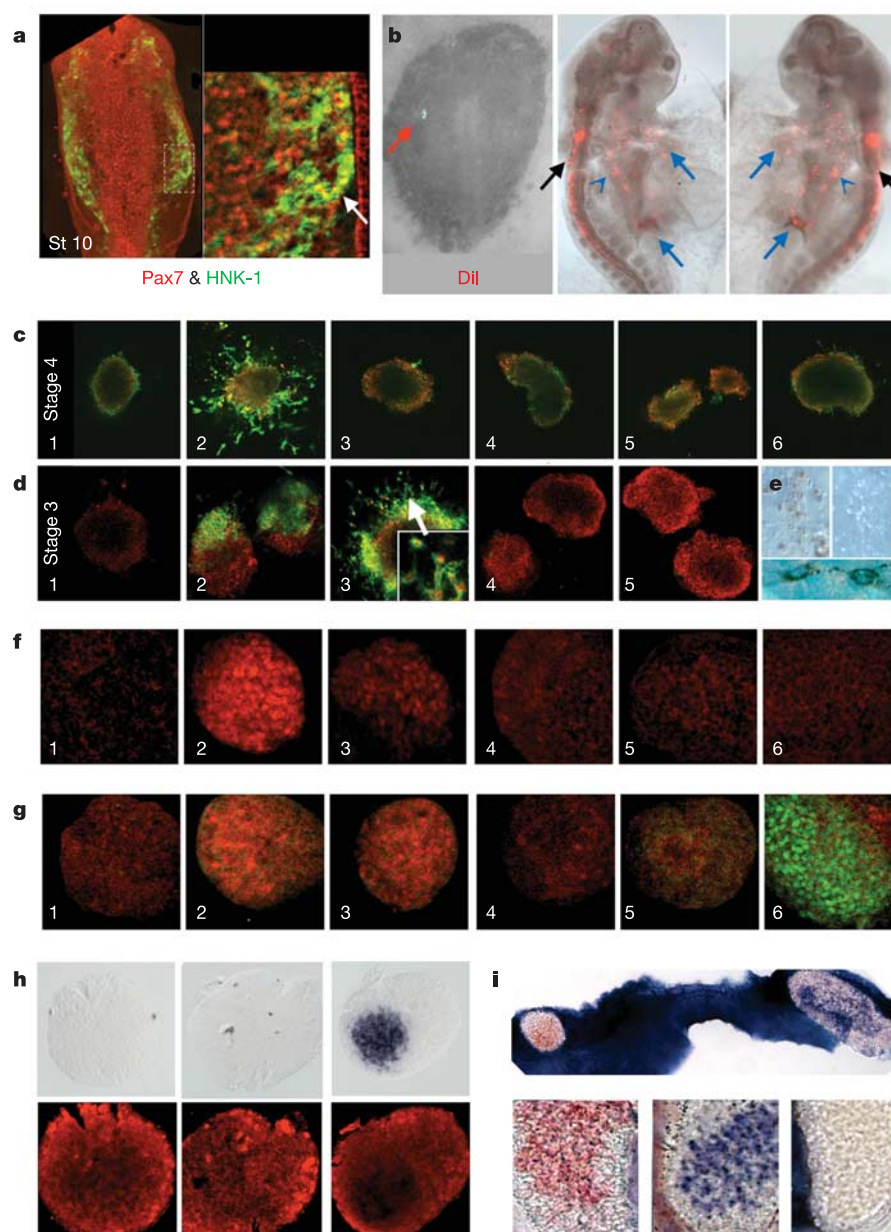


Figure 2 | Neural crest stem cells are specified at stage 3, independently of neural and mesodermal tissues. **a**, Migrating cranial neural crest cells (stage 10 embryo) expressing Pax7 (red) and HNK-1 (green). Right panel shows a higher-magnification confocal image of the framed region, with the white arrow indicating co-localization of Pax7 and HNK-1. **b**, A stage 4 embryo injected with Dil in medial epiblast, before (left) and after (middle and right) 24 h of incubation. Middle and right panels show dorsal and ventral views of the same embryo showing labelled neural folds (black arrows), late migrating crest cells near the neural folds, and stereotypical patterns of neural crest cells migrating to the branchial arches (blue arrowhead), heart and gut (blue arrows). **c–g**, A medial region of the early epiblast is specified to generate neural crest in gastrulating embryos. A strip of epiblast perpendicular to the primitive streak was dissected into 80- μm^2 pieces (sets of tissues from the area opaca/area pellucida border the primitive streak on each side of the primitive streak). Panels **c**, **d**, **f** and **g** show explants from a strip, extending from left to right according to their distance from the primitive streak, from most lateral (1) to most medial (6, primitive streak). **c**, **d**, Explants from stage 4 (**c**) and stage 3 (**d**) embryos cultured for 36–40 h and immunostained for Pax7 (red) and HNK-1 (green). Migratory neural crest cells are visible in explants 2 (**c**) and 2 and 3 (**d**). Inset

in **d** shows a higher-magnification image of HNK1⁺Pax7⁺ migratory crest cells. **e**, Explants cultured for 8 additional days under conditions permissive for melanogenesis contain melanocytes in medial epiblast explants (top left and bottom) but not from adjacent regions (top right). **f–i**, Explants from stage 3 embryos. **f**, Nuclear expression of Pax7 (red) from stage 3 explants is only seen in medial regions after 10 h (explant 2). **g**, Immunostaining for Brachyury (green) and Pax7 (red) expression in stage 3 explants after a 10-h incubation shows no colocalization of the two markers. Brachyury is restricted to primitive-streak-derived explants whereas Pax7 is seen in medial epiblast explants (explants 2–3). **h**, Combined whole-mount *in situ* hybridization for *Tbx6L* (top, blue), and immunofluorescence staining for Pax7 (bottom, red) performed on medial epiblast explants extracted from the collagen after 10 h of incubation. No overlap was observed in most (5/6) medial explants. Top left and middle panels show no overlap, expressing exclusively Pax7. Top right panel shows co-expression of Pax7 and *Tbx6L*. **i**, *In situ* hybridization for *Tbx6L* (blue) and *Pax7* (red) in sections of medial epiblast explants cultured in collagen for 14 h. Top, a section containing one explant positive for Pax7 and one for *Tbx6L*. Bottom, higher-magnification images of three explants showing Pax7 (left), *Tbx6L* (middle) or no expression at all (right).

ectoderm and/or mesoderm⁴. Here we present the finding that discrete regions of epiblast from stage 3–4 chick embryos are already specified to form neural crest. The medial epiblast will express Pax7 and generate neural crest cells in isolation from other tissues. Neural induction is underway by stage 3 but requires later signals that emanate from the node and mesoderm to initiate neural fate²⁹; thus, our explants come from embryos without definitive neural tissue, such that neural tissue cannot account for neural crest formation.

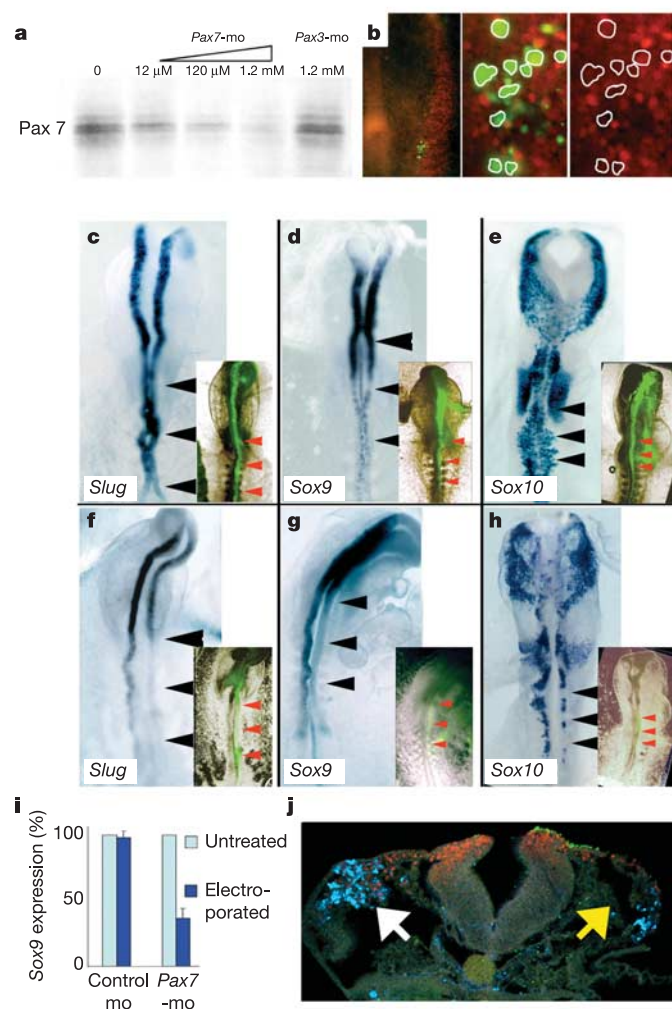


Figure 3 | Pax7 is required for early neural crest formation *in vivo*.

a, *In vitro* translation of Pax7 mRNA showing a reduction in Pax7 protein levels in the presence of increasing concentrations of morpholino against Pax7 (Pax7-mo). Morpholinos against Pax3 (Pax3-mo) do not affect Pax7 translation. **b**, Confocal image of a stage 6 embryo 8 h after electroporation of Pax7-mo (green) and stained with an anti-Pax7 antibody (red). Middle and right panels show higher-magnification images of the region containing morpholinos, showing the cells that received the morpholinos (middle), and the reduction in Pax7 levels in these same cells (right). **c–h**, Pax7-mo (**f, g, h**) downregulates expression of the neural crest markers *Slug*, *Sox9* and *Sox10* (black arrowheads), but control morpholinos (**c, d, e**) have no effect on their expression (arrowheads). Insets show the localization of morpholinos before *in situ* hybridization (red arrowheads). **i**, Quantitative analysis of the effects of control morpholinos (5 base-pair mismatch) and Pax7-mo on *Sox9* expression compared to untreated neural fold (100%). Error bars show s.d. Neural folds electroporated with Pax7-mo show a 65% decrease in *Sox9* transcript levels, and neural folds electroporated with control morpholinos show no significant difference in *Sox9* expression compared to untreated neural folds. **j**, Midbrain section of a stage 10 embryo electroporated with mo-Pax7 (green), showing a reduction in Pax7 (red) and HNK-1 (blue, yellow arrow) levels on the electroporated side. The white arrow points to the normal HNK-1 staining on the untreated side.

Furthermore, we show that the medial epiblast is specified to form neural crest in the absence of definitive neural and early mesodermal markers. Thus, the mechanisms underlying the ability to recapitulate neural crest formation *in vitro* at neurula stages by juxtaposing neural tissue with non-neural ectoderm or mesoderm may be different from those functioning *in vivo* in the early gastrula. Our results indicate that formal establishment of a neural plate border might not be requisite for neural crest specification or induction, and that border formation and neural crest induction might be separable events. This is consistent with the observation that overexpression of *Dlx5* creates an ectopic neural plate border but does not induce neural crest³⁰. Together, these results indicate that all necessary components for generating neural crest cells might be present in the medial epiblast of gastrulating avian embryos.

Our results place the initiation of neural crest induction at or before gastrulation. We establish Pax7 as an early marker required for neural crest formation in avian embryos. In contrast with previous models, we find that neither mesoderm nor proper neural tissues are required for early neural crest induction.

METHODS

Details of *in situ* hybridization and immunohistochemistry, morpholino electroporation, *in vitro* translation, embryo dissection and tissue culture, and genes, primers and product sizes for RT-PCR are given in Supplementary Information.

Received 10 November 2005; accepted 23 February 2006.

- His, W. *Untersuchungen über die erste Anlage des Wirbeltierleibes. Die erste Entwicklung des Hühnchens im Ei* (Vogel, Leipzig, 1868).
- Le Douarin, N. & Kalcheim, C. *The Neural Crest* (Cambridge Univ. Press, Cambridge, 1999).
- Wilkie, A. O. & Morriss-Kay, G. M. Genetics of craniofacial development and malformation. *Nature Rev. Genet.* **2**, 458–468 (2001).
- Basch, M. L., Garcia-Castro, M. I. & Bronner-Fraser, M. Molecular mechanisms of neural crest induction. *Birth Defects Res. C Embryo Today* **72**, 109–123 (2004).
- Moury, J. D. & Jacobson, A. G. Neural fold formation at newly created boundaries between neural plate and epidermis in the axolotl. *Dev. Biol.* **133**, 44–57 (1989).
- Selleck, M. A. & Bronner-Fraser, M. Origins of the avian neural crest: the role of neural plate-epidermal interactions. *Development* **121**, 525–538 (1995).
- Dickinson, M. E., Selleck, M. A., McMahon, A. P. & Bronner-Fraser, M. Dorsalization of the neural tube by the non-neural ectoderm. *Development* **121**, 2099–2106 (1995).
- Basler, K., Edlund, T., Jessell, T. & Yamada, T. Control of cell pattern in the neural tube: regulation of cell differentiation by *dorsalin-1*, a novel TGF β family member. *Cell* **73**, 687–702 (1993).
- Liem, K. F. Jr, Tremml, G., Roelink, H. & Jessell, T. M. Dorsal differentiation of neural plate cells induced by BMP-mediated signals from epidermal ectoderm. *Cell* **82**, 969–979 (1995).
- García-Castro, M. I., Marcelle, C. & Bronner-Fraser, M. Ectodermal Wnt function as a neural crest inducer. *Science* **297**, 848–851 (2002).
- Saint-Jeannet, J. P., He, X., Varmus, H. E. & Dawid, I. B. Regulation of dorsal fate in the neuraxis by *Wnt-1* and *Wnt-3a*. *Proc. Natl Acad. Sci. USA* **94**, 13713–13718 (1997).
- LaBonne, C. & Bronner-Fraser, M. Neural crest induction in *Xenopus*: Evidence for a two-signal model. *Development* **125**, 2403–2414 (1998).
- Mayor, R., Guerrero, I. & Martinez, C. Role of FGF and noggin in neural crest induction. *Dev. Biol.* **189**, 1–12 (1997).
- Basch, M. L., Selleck, M. A. & Bronner-Fraser, M. Timing and competence of neural crest formation. *Dev. Neurosci.* **22**, 217–227 (2000).
- Chi, N. & Epstein, J. A. Getting your Pax straight: Pax proteins in development and disease. *Trends Genet.* **18**, 41–47 (2002).
- Nieto, M. A., Sargent, M. G., Wilkinson, D. G. & Cooke, J. Control of cell behaviour during vertebrate development by *Slug*, a zinc finger gene. *Science* **264**, 835–839 (1994).
- Cheung, M. & Briscoe, J. Neural crest development is regulated by the transcription factor *Sox9*. *Development* **130**, 5881–5893 (2003).
- Streit, A. et al. Chordin regulates primitive streak development and the stability of induced neural cells, but is not sufficient for neural induction in the chick embryo. *Development* **125**, 507–519 (1998).
- Streit, A. & Stern, C. D. Establishment and maintenance of the border of the neural plate in the chick: involvement of FGF and BMP activity. *Mech. Dev.* **82**, 51–66 (1999).
- Pera, E., Stein, S. & Kessel, M. Ectodermal patterning in the avian embryo: epidermis versus neural plate. *Development* **126**, 63–73 (1999).

21. Bang, A. G., Papalopulu, N., Kintner, C. & Goulding, M. D. Expression of *Pax-3* is initiated in the early neural plate by posteriorizing signals produced by the organizer and by posterior non-axial mesoderm. *Development* **124**, 2075–2085 (1997).
22. Garcia-Martinez, V., Alvarez, I. S. & Schoenwolf, G. C. Locations of the ectodermal and nonectodermal subdivisions of the epiblast at stages 3 and 4 of avian gastrulation and neurulation. *J. Exp. Zool.* **267**, 431–446 (1993).
23. Fernández-Garre, P., Rodríguez-Gallardo, L., Gallego-Díaz, V., Alvarez, I. S. & Puellas, L. Fate map of the chicken neural plate at stage 4. *Development* **129**, 2807–2822 (2002).
24. Goulding, M. D., Chalepakis, G., Deutsch, U., Erselius, J. R. & Gruss, P. *Pax-3*, a novel murine DNA binding protein expressed during early neurogenesis. *EMBO J.* **10**, 1135–1147 (1991).
25. Mansouri, A., Stoykova, A., Torres, M. & Gruss, P. Dysgenesis of cephalic neural crest derivatives in *Pax7*^{−/−} mutant mice. *Development* **122**, 831–838 (1996).
26. Mansouri, A. The role of *Pax3* and *Pax7* in development and cancer. *Crit. Rev. Oncog.* **9**, 141–149 (1998).
27. Mansouri, A. & Gruss, P. *Pax3* and *Pax7* are expressed in commissural neurons and restrict ventral neuronal identity in the spinal cord. *Mech. Dev.* **78**, 171–178 (1998).
28. Monsoro-Burq, A. H., Wang, E. & Harland, R. *Msx1* and *Pax3* cooperate to mediate FGF8 and WNT signals during *Xenopus* neural crest induction. *Dev. Cell* **8**, 167–178 (2005).
29. McLarren, K. W., Litsiou, A. & Streit, A. *DLX5* positions the neural crest and preplacode region at the border of the neural plate. *Dev. Biol.* **259**, 34–47 (2003).
30. Sheng, G., dos Reis, M. & Stern, C. D. Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell* **115**, 603–613 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Streit, S. Fraser, A. Groves, L. Gammill, V. Lee, J. Sechrist and T. Moreno for comments and discussions; S. Mackem, D. Wilkinson, E. Matsunaga, A. Kawakami, O. Pourquie and H. Kondo for gifts of *Tbx6l*, *cMsx1*, full-length *cPax3*, full-length *cPax7*, *BMP4*, *Sox2*, *Dlx5* and *Gata2* probes and constructs; and D. Arce and Y. Liu for technical assistance. This work was supported by a USPHS grant to M.B.-F.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.I.G.-C. (martin.garcia-castro@yale.edu).

Neurons in the orbitofrontal cortex encode economic value

Camillo Padoa-Schioppa¹ & John A. Assad¹

Economic choice is the behaviour observed when individuals select one among many available options. There is no intrinsically 'correct' answer: economic choice depends on subjective preferences. This behaviour is traditionally the object of economic analysis¹ and is also of primary interest in psychology². However, the underlying mental processes and neuronal mechanisms are not well understood. Theories of human and animal choice^{1–3} have a cornerstone in the concept of 'value'. Consider, for example, a monkey offered one raisin versus one piece of apple: behavioural evidence suggests that the animal chooses by assigning values to the two options⁴. But where and how values are represented in the brain is unclear. Here we show that, during economic choice, neurons in the orbitofrontal cortex^{5–18} (OFC) encode the value of offered and chosen goods. Notably, OFC neurons encode value independently of visuospatial factors and motor responses. If a monkey chooses between A and B, neurons in the OFC encode the value of the two goods independently of whether A is presented on the right and B on the left, or vice versa. This trait distinguishes the OFC from other brain areas in which value modulates activity related to sensory or motor processes^{19–25}. Our results have broad implications for possible psychological models, suggesting that economic choice is essentially choice between goods rather than choice between actions. In this framework, neurons in the OFC seem to be a good candidate network for value assignment underlying economic choice.

In our experiments, monkeys choose between two types of juice (A and B; where A is preferred) offered in different amounts. For example, in the session shown in Fig. 1, the monkey chooses between water (juice A) and unsweetened Kool-Aid (juice B). Offer types include 1B:2A, 1B:1A, 2B:1A, 3B:1A, 4B:1A, 6B:1A, and 10B:1A, and the 'forced choices' 0B:1A and 3B:0A. Behaviourally, we observe a trade-off between juice type and juice quantity. The monkey chooses A when 1B, 2B, or 3B are available as alternatives, it is roughly indifferent between the two juices when offered 4B:1A, and it chooses B when 6B or 10B are available.

We interpret this pattern of choice in terms of the 'relative value' of the two juices⁴: in this case, the value of 1A is roughly equal to the value of 4B. Fitting a sigmoid curve provides the better estimate $V(1A) = V(4.1B)$, where $V(x)$ indicates the value of x . Assuming a linear value function, we obtain $V(A) = 4.1V(B)$. This equation puts different quantities of juices A and B on the same value scale. On this basis, we can compute for each trial the value of the juice chosen by the monkey. Expressing values in units of $V(B)$, the variable 'chosen value' is about 4 when the monkey chooses 1A or 4B. When the monkey chooses 2A, the chosen value is about 8. When the monkey chooses 6B, 10B or 3B, the chosen value is respectively equal to 6, 10 or 3. Hence, we can make specific hypotheses regarding the neuronal representation of juice values. In different sessions and with different juices, we record different behavioural choice patterns. We then

analyse each cell in relation to the choice pattern recorded in the same session.

Our recordings focused on area 13 in the OFC. Figure 2 illustrates the activity of one representative neuron. The cell's activity does not depend on whether juice A is offered on the left or on the right (Fig. 2b). It also does not depend on whether the monkey chooses the juice on the left or the juice on the right (that is, makes an eye movement to the left or to the right; Fig. 2c). However, the cell's activity varies with the offer type. This is consistent across the neuronal population. We recorded the activity of 931 cells and we analysed their neuronal responses in seven time windows (see Methods). We tested the activity of each cell in each time window with a three-way analysis of variance (ANOVA, with factors: [position of juice A] $\times$ [movement direction] $\times$ [offer type], $P < 0.001$). Rarely do responses depend on either the spatial configuration of the offers or the motor output ($<5\%$ neurons). In contrast, the activity of 505 (54%) neurons varies significantly depending on the offer type in at least one time window. Pooling time windows, a total of 1,379 responses are significantly modulated by the offer type (Supplementary Fig. S2).

The cell shown in Fig. 2d has a U-shaped response similar to that hypothesized for a neuron encoding the *chosen value*. For this session $V(A) = 1.9V(B)$. Accordingly, the activity of the cell is

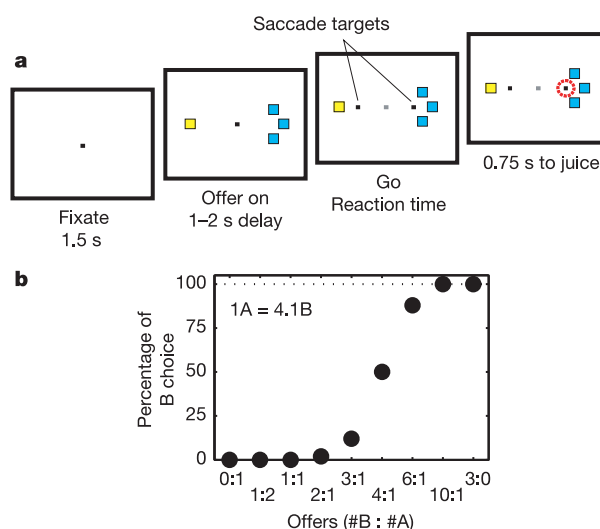


Figure 1 | Experimental design. **a**, Trial structure (see Methods). **b**, Example of behavioural choice pattern. The plot shows the percentage of trials in which the monkeys chose juice B (y axis) for various offer types (x axis). A sigmoid fit provides the measure of the relative value $n^* = 4.1$. The dotted red circle indicates the saccade target chosen by the monkey.

¹Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA.

low when the monkey chooses 1A or 2B (in units of $V(B)$, *chosen value* ≈ 2), it is higher when the monkey chooses 2A or 4B (*chosen value* ≈ 4), and it is highest when the monkey chooses 3A or 6B (*chosen value* ≈ 6). A linear regression of this response on the variable *chosen value* provides $R^2 = 0.86$. Similar U-shaped responses are frequent in the OFC, and Fig. 3a–c illustrates three more examples. We also find other types of responses. For example, neuronal responses often reflect the value of one of the two juices alone. Figure 3d, e shows two cells in which activity covaries with the value of A offered and the value of B offered, respectively. We label these responses as related to the variable ‘offer value’. Other frequently observed responses vary in a binary fashion depending on the type of juice chosen by the monkey, independently of the amount (Fig. 3f). We interpret these responses as related to the variable juice ‘taste’.

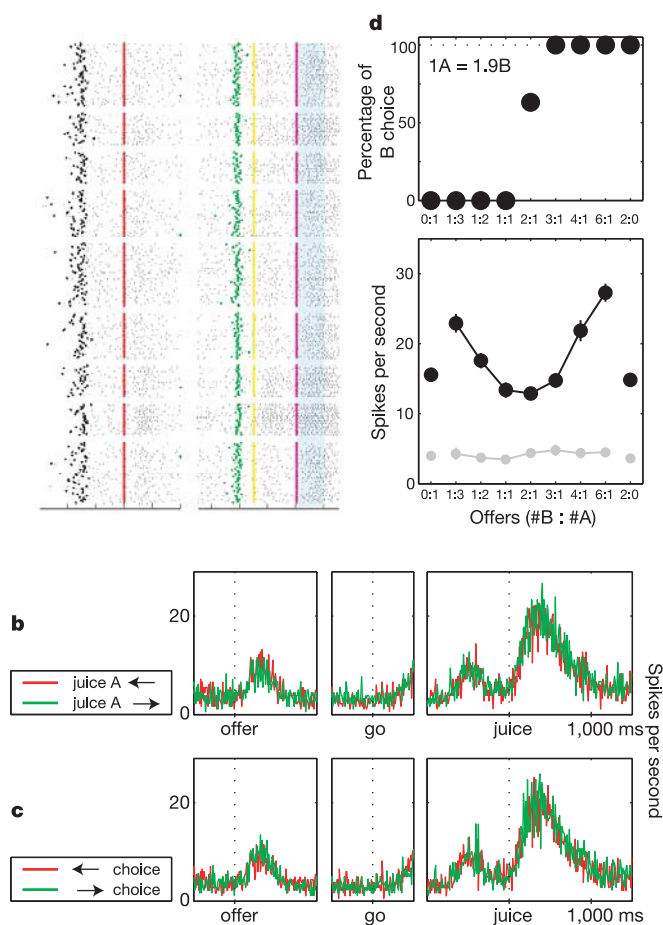


Figure 2 | Activity of one neuron. **a**, Rasters. Each line represents one trial and each small dot represents one spike. Trials, arranged by offer type, are aligned at the ‘offer’ (left) and at the ‘juice’ (right). The blue highlight marks the post-juice time window. ‘Sacc’ indicates the time of the saccade. **b**, Activity profiles shown separately for trials in which juice A is offered on the left (red) or on the right (green). The cell activity does not depend on the spatial configuration of the visual stimulus. **c**, Activity profiles shown separately for trials in which the monkey chooses the juice offered on the left (red) or on the right (green). The cell activity does not depend on the direction of the eye movement. **d**, The top panel shows the choice pattern recorded in this session ($n^* = 1.9$). The bottom panel shows the activity of the cell ($\pm$ s.e.m.) recorded in the pre-offer (light grey, control) and post-juice (black) time windows. Note that the response does not reflect simple physical properties of the visual stimulus, such as the number of squares displayed on the monitor. For example, offer types 1B:3A and 3B:1A, which are visually identical except for the colour of the squares, elicit very different activation.

Although many neurons seem to encode *chosen value*, *offer value* or juice *taste*, the relation between their activity and these three variables could be subordinate to a correlation with other behavioural variables. For example, neurons in the OFC might encode the number of squares on the monitor (or variables proportional to number, such as juice quantity, or absolute luminance of the visual stimulus). To cast a wide net, we examine the linear dependence of neuronal data on 19 possible variables (Supplementary Fig. S1). For example, we analyse the variables ‘*chosen number*’, ‘*total number*’ and ‘*total value*’. We include in this analysis 1,379 responses significantly modulated by the offer type, and we regress each response separately on each variable. Collectively, the 19 variables explain 1,227 (89%) neuronal responses. However, the 19 variables are often highly correlated.

To identify a few variables that best describe the neuronal population, we adapt procedures for variable selection commonly used in multilinear regression in the presence of multi-collinearity. Both the stepwise and the best-subset methods identify the variables *offer value*, *chosen value* and *taste*, which explain well the large majority of responses (1,085/1,379 = 79% responses, with mean $R^2 = 0.63$). A post-hoc analysis indicates that the explanatory power of these three variables is significantly higher than that of challenging alternatives. Furthermore, data from the two monkeys analysed separately provide statistically indistinguishable results. Finally, a bilinear regression analysis indicates that in 890/1,085 (82%) cases, adding a second variable or a quadratic value term does not improve the regression significantly (Supplementary Results S5 to S10 and Supplementary Figs S4 to S11). We conclude that, as a population, OFC responses indeed encode the variables *offer value*, *chosen value* and *taste*.

We next turn to a specific analysis of U-shaped responses (Figs 2d, 3a–c). In our experiments, relative values were generally stable within any recording session. However, the relative value of any given pair of juices could vary from day to day. For example, the relative value of apple juice versus peppermint tea varied between 1.5 and 3. This variability provides a further opportunity to test the neuronal encoding of value; specifically, U-shaped responses should reflect this variability. For this analysis, we test the entire neuronal population with the regression function $a_0 + a_A(m_A) + a_B(m_B)$, where m_A and m_B represent the amounts of juices A and B chosen by the monkey. We define a response to be U-shaped if both a_A and a_B differ from zero ($P < 0.01$). If U-shaped responses indeed encode the value of the chosen juice, the slope ratio $k^* = a_A/a_B$ should be, for each U-shaped response, equal to the relative value (n^*) measured in that session. Most importantly, the slope ratio k^* obtained for different responses should covary with n^* . To test this prediction, we compute the regression $k^* = b_0 + b_1 n^*$ separately for every juice pair. Averaging across juice pairs, we obtain b_0 ($\pm$ s.e.m.) = -0.13 (± 0.15) and $b_1 = 1.05$ (± 0.15), consistent with the identity $k^* = n^*$. This result demonstrates that U-shaped responses do not reflect the quantity of any particular juice ingredient (for example, sugar). Rather, they encode the value monkeys assign to the juice they choose to consume (Supplementary Results S10 to S12 and Supplementary Figs S12 to S14).

The timing of neuronal activation is as follows: the average neuronal activity peaks shortly after the offer, declines during the delay, is low before and during the eye movement, and has two secondary peaks at juice delivery (Supplementary Fig. S3). To appreciate how the variables *offer value*, *chosen value* and *taste* are represented in the OFC over time, we analyse the activity of each cell in 50-ms, non-overlapping time bins. Figure 4 shows the number of cells encoding each of the three variables at different times. Remarkably, the time profile of different variables seems to reflect the mental processes the monkey presumably undertakes during a trial. Shortly after the offer, when the monkey assigns values to the two juices, neurons encoding the *offer value* (that is, the value of one juice or the other) are most prevalent. Also during the delay, many neurons

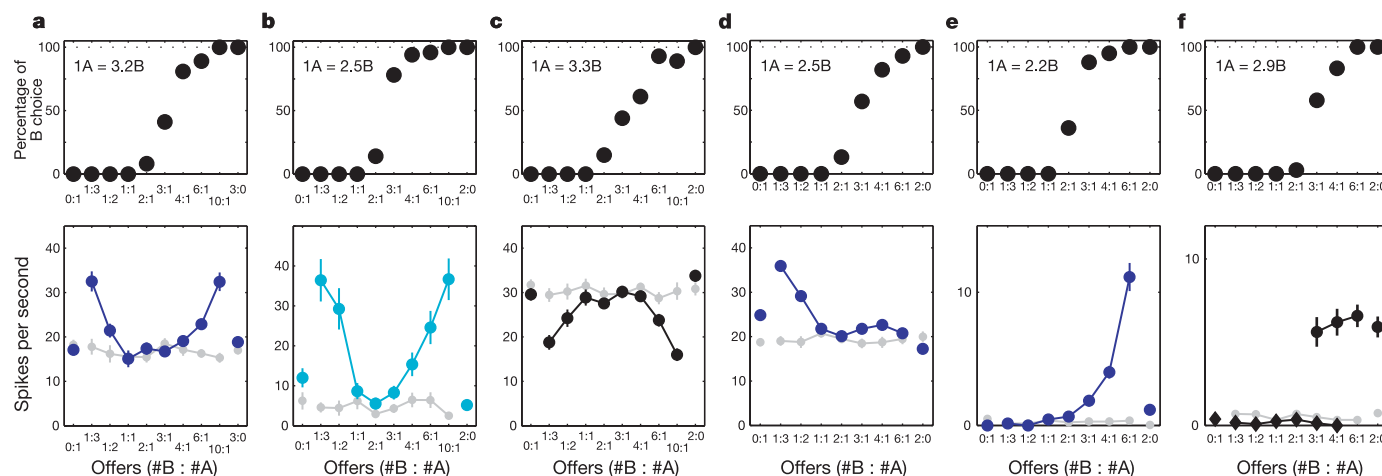


Figure 3 | Activity of six neurons. For each cell, the top panel shows the choice pattern, with the relative value indicated on the top left. The bottom panel shows the activity of the cell. **a–c**, Responses encoding the *chosen value*. The response in **c** is negatively correlated with the *chosen value* (high activity for low value). **d–e**, Responses encoding the value of juice A offered (**d**) and the value of juice B offered (**e**). We refer to these responses as related to the *offer value*. **f**, Response encoding the juice *taste*. Here we separate trials

encode the *chosen value* (that is, the value of the juice the monkey will eventually consume), even though the choice is still covert (because the ‘go’ signal has not been given yet). Finally, after the monkey has indicated its choice, before and after juice delivery, many neurons encode the *taste* of the chosen juice.

Conceptually, responses encoding the *chosen value* are particularly interesting because, in addition to being independent of the visuomotor contingencies of the task, they are also independent of the specifics of the good, namely juice type and juice amount. These responses encode economic value in a non-specific way. Further research is necessary to establish whether this result generalizes to other kinds of goods, such as non-comestible goods²⁶. The interpretation of *offer value* responses is made more cautiously because, assuming linear value functions, the value of a given amount of juice is proportional to the juice quantity.

‘Value’ is known to modulate the activity of neurons in several sensory and motor areas^{19–25}. For example, neurons in the lateral intraparietal area activate when monkeys plan a saccade towards a particular location of the visual field; their response is enhanced

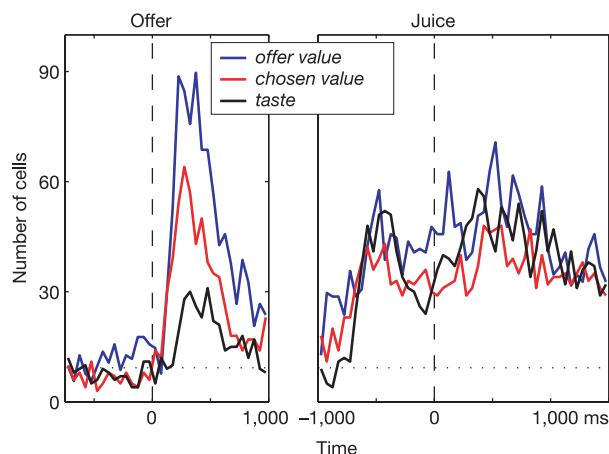


Figure 4 | Time course. We assign each neuron to one of the three variables only if the regression slope is significantly different from zero ($P < 0.01$), and we include all 931 neurons in the analysis. The dotted line indicates chance level (9.31).

in which the monkey chose juice A (diamonds) or juice B (circles). The response reflects the chosen juice type independently of the amount. Responses were recorded in the post-offer (**a, d, e**, blue), pre-juice (**b**, cyan), and post-juice (**c, f**, black) time windows. For each cell, the curves in light grey show the activity in the pre-offer time window. Error bars represent s.e.m.

when the eye movement is associated with higher value²⁰. On this basis, it has been proposed that parietal neurons encoding the value of all possible courses of action form a common path for decision-making, and that their activity is actually the subject of economic theory²⁷. According to this ‘action-based’ model, economic choice is fundamentally choice between actions. That neurons in the OFC encode the economic value of offered and chosen goods *per se*, as opposed to reflecting value as a modulation of visuomotor processes, suggests an alternative ‘good-based’ model, according to which economic choice is fundamentally choice between goods. In this view, choice is made between goods, and a suitable motor action is subsequently planned and executed.

Several arguments seem to favour the good-based model. From a computational perspective, a modular design separating the mental operations of ‘choosing’ and ‘moving’ is more parsimonious^{28,29}. In addition, values processed in the OFC are logically sufficient for good-based choice. The action-based model would thus imply that, during economic choice, the nervous system operates in a computationally inefficient way, while undertaking all the processes needed to choose efficiently. Finally, a vast literature links choice in various domains to the OFC. For example, human patients and monkeys with OFC lesions can present eating disorders and hyperorality^{6–8}, abnormal risk-seeking and gambling behaviour^{9,10}, and impulsivity, altered personality and abnormal social behaviour^{6,11}. In contrast, parietal lesions typically result in visuospatial deficits such as hemi-neglect or Balint’s syndrome³⁰. In conclusion, together with other lines of evidence, the present results support a good-based psychological model of economic choice behaviour.

METHODS

Each trial begins with the monkey fixating the centre of a computer monitor (Fig. 1a). After 1.5 s, two sets of squares appear on opposite sides of the fixation point (‘offer’). The colour of the squares indicates the juice type and the number of squares indicates the juice amount. For example, a monkey offered three blue squares versus one yellow square chooses between three drops of peppermint tea and one drop of grape juice. After a randomly variable delay (1–2 s), two saccade targets appear near the offers (‘go’). The monkey indicates its choice with an eye movement, and must maintain fixation on the target for an additional 0.75 s before juice delivery (‘juice’). The trial is aborted if the monkey breaks fixation before the go signal. The amounts of the two offered juices (0–10 drops) vary pseudo-randomly. For a given offer type, left/right positions are counter-balanced (that is, the monkey may be offered 1A on the left and 3B on the

right, or vice versa). A variety of different pairs of juices are used in different sessions (Supplementary Methods).

We analyse cell activity in the following time windows: 0.5 s pre-offer (a control time window); 0.5 s post-offer; late delay (0.5–1.0 s after the offer); 0.5 s pre-go; reaction time (from go to saccade); 0.5 s pre-juice; and 0.5 s post-juice. For the statistical analysis, we separate for each offer type trials in which the monkey chooses juices A and B (Supplementary Methods).

Received 28 November 2005; accepted 24 February 2006.

Published online 23 April 2006.

- Samuelson, P. A. *Foundations of Economic Analysis* (Harvard Univ. Press, Cambridge, 1947).
- Kahneman, D. & Tversky, A. (eds) *Choices, Values and Frames* (Russell Sage Foundation–Cambridge Univ. Press, Cambridge/New York, 2000).
- Kagel, J. H., Battalio, R. C. & Green, L. *Economic Choice Theory: An Experimental Analysis of Animal Behavior* (Cambridge Univ. Press, Cambridge/New York, 1995).
- Padoa-Schioppa, C., Jandolo, L. & Visalberghi, E. Multi-stage mental process for economic choice in capuchins. *Cognition* **99**, B1–B13 (2006).
- Ongur, D. & Price, J. L. The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cereb. Cortex* **10**, 206–219 (2000).
- Pasquier, F. & Petit, H. Frontotemporal dementia: its rediscovery. *Eur. Neurol.* **38**, 1–6 (1997).
- Hodges, J. R. Frontotemporal dementia (Pick's disease): clinical features and assessment. *Neurology* **56**, S6–S10 (2001).
- Butter, C. M., McDonald, J. A. & Snyder, D. R. Orality, preference behavior, and reinforcement value of nonfood object in monkeys with orbital frontal lesions. *Science* **164**, 1306–1307 (1969).
- Rahman, S., Sahakian, B. J., Hodges, J. R., Rogers, R. D. & Robbins, T. W. Specific cognitive deficits in mild frontal variant frontotemporal dementia. *Brain* **122**, 1469–1493 (1999).
- Bechara, A., Tranel, D., Damasio, H. & Damasio, A. R. Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cereb. Cortex* **6**, 215–225 (1996).
- Raleigh, M. J. & Steklis, H. D. Effect of orbitofrontal and temporal neocortical lesions of the affiliative behavior of vervet monkeys (*Cercopithecus aethiops sabaeus*). *Exp. Neurol.* **73**, 378–389 (1981).
- Izquierdo, A., Suda, R. K. & Murray, E. A. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *J. Neurosci.* **24**, 7540–7548 (2004).
- O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neurosci.* **4**, 95–102 (2001).
- Knutson, B., Taylor, J., Kaufman, M., Peterson, R. & Glover, G. Distributed neural representation of expected value. *J. Neurosci.* **25**, 4806–4812 (2005).
- Rolls, E. T., Sienkiewicz, Z. J. & Yaxley, S. Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *Eur. J. Neurosci.* **1**, 53–60 (1989).
- Roesch, M. R. & Olson, C. R. Neuronal activity related to reward value and motivation in primate frontal cortex. *Science* **304**, 307–310 (2004).
- Wallis, J. D. & Miller, E. K. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *Eur. J. Neurosci.* **18**, 2069–2081 (2003).
- Tremblay, L. & Schultz, W. Relative reward preference in primate orbitofrontal cortex. *Nature* **398**, 704–708 (1999).
- Kawagoe, R., Takikawa, Y. & Hikosaka, O. Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neurosci.* **1**, 411–416 (1998).
- Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
- Leon, M. I. & Shadlen, M. N. Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron* **24**, 415–425 (1999).
- Shidara, M. & Richmond, B. J. Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science* **296**, 1709–1711 (2002).
- Ikeda, T. & Hikosaka, O. Reward-dependent gain and bias of visual responses in primate superior colliculus. *Neuron* **39**, 693–700 (2003).
- Roesch, M. R. & Olson, C. R. Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *J. Neurophysiol.* **90**, 1766–1789 (2003).
- McCoy, A. N., Crowley, J. C., Haghighian, G., Dean, H. L. & Platt, M. L. Saccade reward signals in posterior cingulate cortex. *Neuron* **40**, 1031–1040 (2003).
- Deaner, R. O., Khera, A. V. & Platt, M. L. Monkeys pay per view: adaptive valuation of social images by rhesus macaques. *Curr. Biol.* **15**, 243–248 (2005).
- Glimcher, P. W., Dorris, M. C. & Bayer, H. M. Physiological utility theory and the neuroeconomics of choice. *Games Econ. Behav.* **52**, 213–256 (2005).
- Simon, H. A. The architecture of complexity. *Proc. Am. Phil. Soc.* **106**, 467–482 (1962).
- Pinker, S. *How the Mind Works* (Norton, New York, 1997).
- Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L. & Squire, L. R. *Fundamental Neuroscience* Ch. 53 (Academic, San Diego, 1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Irwin, T. LaFratta, D. Averbuch, J. LeBlanc, S. Peled and J. Harper for technical assistance and animal care; E. Brown for a discussion on the statistical analysis; and T. Herrington, D. Freedman and E. Bizzi for comments on the manuscript. This work was supported by postdoctoral fellowships from the Lefler Foundation and from the Harvard Mind/Brain/Behavior Initiative (to C.P.-S.) and by a grant from the National Institute of Neurological Disorders and Stroke (to J.A.A.).

Author Contributions C.P.-S. performed all aspects of the study, including the design of the experiment, collecting and analysing the data, and writing the manuscript. J.A.A. assisted in experimental design, data analysis and manuscript preparation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.P.-S. (camillo@alum.mit.edu).

A trehalose metabolic enzyme controls inflorescence architecture in maize

Namiko Satoh-Nagasawa¹, Nobuhiro Nagasawa², Simon Malcomber^{3†}, Hajime Sakai² & David Jackson¹

Inflorescence branching is a major yield trait in crop plants controlled by the developmental fate of axillary shoot meristems¹. Variations in branching patterns lead to diversity in flower-bearing architectures (inflorescences) and affect crop yield by influencing seed number or harvesting ability^{2,3}. Several growth regulators such as auxins, cytokinins and carotenoid derivatives regulate branching architectures⁴. Inflorescence branching in maize is regulated by three *RAMOSA* genes⁵. Here we show that one of these genes, *RAMOSA3* (*RA3*), encodes a trehalose-6-phosphate phosphatase expressed in discrete domains subtending axillary inflorescence meristems. Genetic and molecular data indicate that *RA3* functions through the predicted transcriptional regulator *RAMOSA1* (*RA1*)⁵. We propose that *RA3* regulates inflorescence branching by modification of a sugar signal that moves into axillary meristems. Alternatively, the fact that *RA3* acts upstream of *RA1* supports a hypothesis that *RA3* itself may have a transcriptional regulatory function.

Trehalose is a disaccharide composed of two glucose units. It is present in all kingdoms and has functions in carbohydrate storage, stress protection and metabolic regulation^{6,7}. Until fairly recently it was thought to be present in only a small number of desiccation-tolerant plants (reviewed in ref. 8) because endogenous trehalose levels are very low in most plants. However, trehalose biosynthesis genes are present in all plants, and recent data indicate that this sugar functions in stress protection as well as in carbohydrate utilization and growth^{9–11}. Trehalose biosynthesis occurs in two steps⁸. First, trehalose 6-phosphate (T6P) is made from UDP-glucose and glucose 6-phosphate by T6P synthase (TPS), and T6P is then converted to trehalose by T6P phosphatase (TPP)¹². Here we show that *RAMOSA3* controls maize inflorescence architecture and encodes a functional TPP enzyme, implying a previously unrecognized role for trehalose metabolism in developmental signalling and morphogenesis. Although trehalose biosynthetic enzymes are ubiquitous, this is conclusive evidence that they have a defined developmental function. Our findings greatly extend previous studies that suggested such a function but did not identify any specific developmental process or stage that is affected by such enzymes. For example, heterologous expression of bacterial trehalose enzymes can perturb development^{11,13}. Furthermore, *tps1* mutants in *Arabidopsis* are embryo lethal¹⁰, and *TPS1* is required for sustained growth in *Arabidopsis*, including during the floral transition¹⁴.

ramosa3 (*ra3*) is a classical mutant of maize¹⁵. Maize has two types of flower-bearing structure with different architectures that were selected during maize domestication to enhance its utility as an agricultural crop⁵. The terminal male inflorescence, or tassel, has long branches at its base and a central spike that bears shorter branches containing spikelet pairs (Fig. 1c). In contrast, the female inflorescences, or ears, are positioned laterally and contain only

short branches, a trait that is thought to aid in the efficient packing and harvesting of seeds (Fig. 1a). *RA3* is required for this specialized architecture, because *ra3* mutant tassels have additional long branches (Fig. 1d; wild-type (B73) 7.9 ± 0.2 (mean $\pm$ s.e.m.); *ra3* 15.4 ± 0.8 branches) and *ra3* mutant ears have abnormal long branches at their bases (Fig. 1b). A detailed analysis of ear development by scanning electron microscopy (SEM) showed that there was no morphological difference between wild-type and *ra3* inflorescences

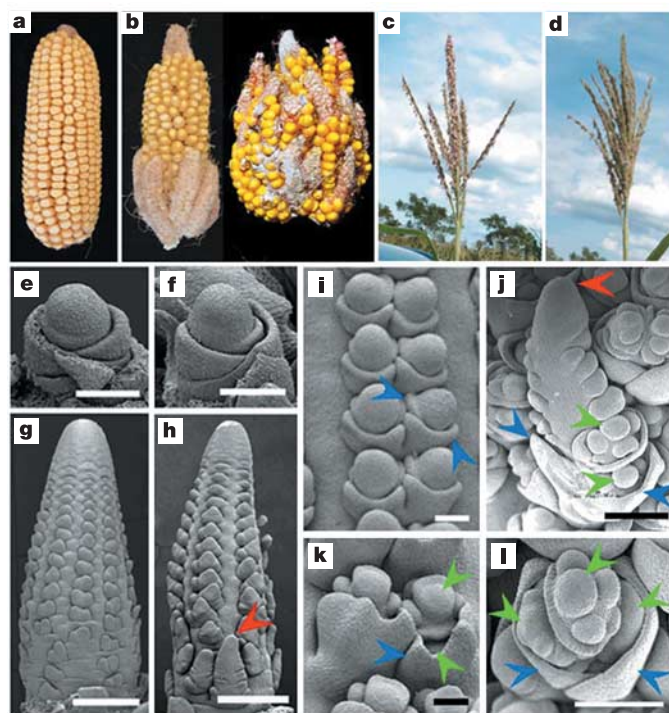


Figure 1 | *ra3* mutant phenotypes. **a**, Mature wild-type ear. **b**, Mature *ra3* ears introgressed into B73 (left) or in a mixed genetic background (right) had abnormal branches and irregular seed rows. **c**, Mature wild-type tassel. **d**, Mature *ra3* tassel with additional long branches. **e–l**, SEM of wild-type ear development (**e**, **g**, **i**, **k**) and *ra3* ear development (**f**, **h**, **j**, **l**): before initiation of axillary meristems (**e**, **f**), ears 2 mm long (**g**, **h**), ears 5 mm long (**i**, **j**) and ears 1 cm long (**k**, **l**). In the *ra3* mutant (**h**) some SPMs changed their identity and formed indeterminate branches, resembling the long branches at the base of the tassel (red arrowhead). SMs in the wild type (**i**, **k**) produced a pair of FMs (green arrowheads) subtended by glumes (blue arrowheads), but in *ra3* (**j**) after the production of FMs they sometimes converted to indeterminate branches. In other cases, SMs in *ra3* (**l**) made multiple FMs. Scale bars, 100 μ m (**e**, **f**, **i**, **j**), 500 μ m (**g**, **h**) and 200 μ m (**k**, **l**).

¹Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. ²DuPont Crop Genetics Experimental Station E353, Wilmington, Delaware 19880, USA. ³Department of Biology, University of Missouri–St Louis, St Louis, Missouri 63121, USA. [†]Present address: Department of Biological Sciences, California State University–Long Beach, Long Beach, California 90840, USA.

before axillary inflorescence meristems were produced (Fig. 1e, f). However, once axillary meristems were initiated in *ra3* mutants, they showed a general loss of determinacy as well as changes in identity (Fig. 1h, j, l). In normal ears, the inflorescence meristem initiates axillary meristems (spikelet pair meristems (SPMs); Fig. 1g), which are determinate structures producing two spikelet meristems (SMs; Fig. 1i). Each SM in turn produces two floral meristems³ (FMs; Fig. 1k). In *ra3* mutants the axillary meristems were enlarged and acquired abnormal identity (Fig. 1h, j) or became indeterminate (Fig. 1l), leading to the production of long branches (Fig. 1h, j) or more FMs (Fig. 1l). Similar developmental defects were observed in the tassel, although at a lower frequency (not shown). We therefore conclude that *RA3* acts to establish the correct identity and determinacy of axillary meristems in both male and female inflorescences.

With the use of bulked segregant analysis¹⁶ *RA3* was mapped to chromosome 7, and by fine mapping it was located to a single bacterial artificial chromosome (BAC), which was sequenced. This sequence was used to design further markers that delimited the *RA3* locus to a single predicted open reading frame (ORF) (Supplementary Fig. 1a, b and Supplementary Table 1). Identification of lesions in seven *ra3* alleles all mapping to the same ORF confirmed unambiguously that this encodes *RA3* (Supplementary Table 2). *RA3* encodes a predicted protein of 361 amino-acid residues with significant similarity to TPPs⁹. A non-conserved amino-terminal region of about 80 residues is followed by the TPP domain, which contains two conserved 'phosphatase boxes'^{8,17} (Supplementary Fig. 1b). Most of the *ra3* mutant alleles encode frame shifts leading to a stop codon before the second phosphatase box and to a strong mutant phenotype. One of them, *ra3-fea1*, has no detectable transcript in immature ears (Supplementary Fig. 1c). They are therefore likely to be null alleles. The *ra3-NI* allele has a milder phenotype (not shown) and has a premature stop after the second phosphatase box (Supplementary Table 2).

About 10 kilobases upstream of *RA3*, a highly similar gene was discovered and dubbed *SISTER OF RA3* (*SRA*) (Supplementary Fig. 1a). The conserved syntenic region of rice contained only a single TPP gene, gi33146623. *RA3*, *SRA* and gi33146623 share about 70% amino-acid identity within the TPP domain, and about 65% identity overall. We searched for closely related homologues in GenBank and in the maize genome sequence assemblies^{18–20} and amplified *RA3/SRA*-like genes from a range of other grasses. A phylogenetic analysis of TPP genes was conducted with bayesian methods²¹ (Supplementary Fig. 2). This phylogenetic analysis estimates a well-supported *RA3/SRA* clade (100% clade credibility [CC]) nested within a well-supported clade (97% CC) of grass TPP genes. Although the exact placement of the *RA3/SRA* duplication event is still in question, our best estimate is that it occurred near the base of the major diversification of the grasses. Because of multiple duplication events within or before the origin of the grasses, the closest *Arabidopsis* TPP genes (*TPPB* (ref. 22) and *At1g22210*) cannot be considered orthologous to either *RA3* or *SRA*.

RA3 was expressed predominantly in young inflorescences, at the stage where axillary meristem primordia were being initiated. In contrast, *SRA* was expressed more widely in all organs tested and showed the highest expression in roots (Fig. 2a). *In situ* hybridization revealed a localized pattern of *RA3* expression in cup-shaped domains at the base of axillary meristems in young ear primordia, and in a stripe between upper and lower florets (Fig. 2b–f). This expression was specific for *RA3*, because it was not seen in *ra3-fea1* mutant ears (Fig. 2g). In the tassel, *RA3* was not expressed as the long branches were initiated (not shown) but was expressed at the base of SPMs (Fig. 2h). These patterns are consistent with a function of *RA3* in promoting determinacy of axillary meristems in the tassel. Together with the developmental analysis, the restricted expression pattern suggests a highly specific developmental role for *RA3* in inflorescence development. *In situ* hybridization with an *SRA*-specific

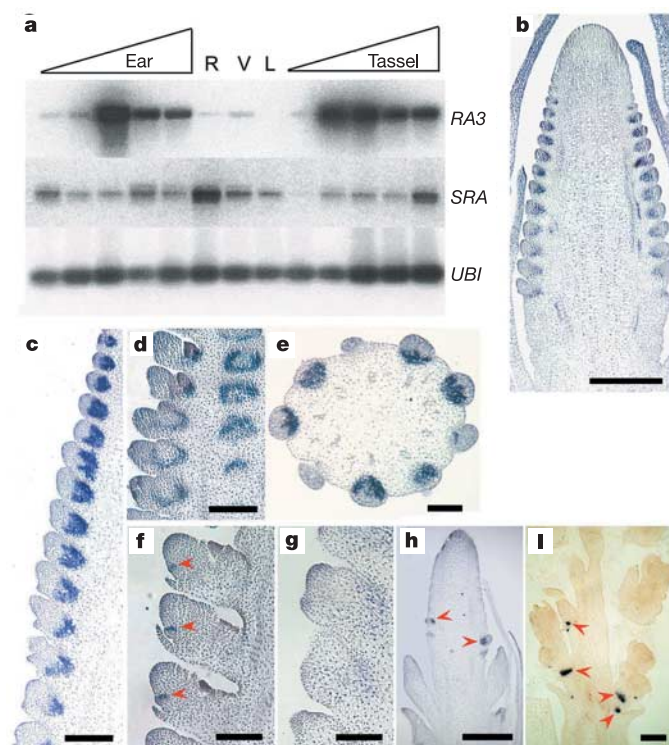


Figure 2 | Developmental expression of *RA3* and *SRA*. **a**, RT-PCR of mRNAs from root (R), vegetative apex (V), young leaves (L) and ear or tassel inflorescence primordia. The triangles represent increasing inflorescence size, from the stage before axillary meristem initiation to ears and tassels about 1.5 cm long. **b–f**, Detection of *RA3* expression by *in situ* hybridization. **b**, Longitudinal section of an ear 2 mm long. **c**, **d**, Longitudinal median and glancing sections of ears 7 mm long. **e**, Transverse section of ear 7 mm long. **f**, Longitudinal section of developing spikelets. **g**, Control section showing lack of hybridization signal in a *ra3-fea1* ear. **h**, Longitudinal section of a tassel 2 mm long. **i**, *In situ* hybridization of the rice *SRA* gene in a longitudinal section of a rice inflorescence. Red arrowheads show the expression domain of *RA3* (**f**, **h**) and rice *SRA* (**i**). Scale bars, 500 μ m (**b**, **h**), 200 μ m (**c**, **d**, **e**) and 100 μ m (**f**, **g**, **i**).

probe in maize inflorescences failed to detect a localized expression pattern, indicating that it might be expressed at a low level in all or many cells (not shown).

To investigate whether *RA3*-like genes might also have a developmental function in other plants, we isolated a *RA3* orthologue from sorghum and examined its expression pattern. This gene was expressed in a similar localized manner to that of *RA3* (not shown), suggesting that *RA3* function is conserved in other grasses. We also examined expression of the closest rice homologue, which was phylogenetically most similar to *SRA* (Supplementary Fig. 2). It was expressed widely, with higher expression levels in both root and young inflorescence (Supplementary Fig. 3). However, like *RA3* it showed a localized expression domain at the base of axillary inflorescence branches and beneath the spikelets (Fig. 2i), indicating that this gene might also regulate inflorescence development in rice.

We next examined whether *RA3* had TPP activity, a possibility indicated by its sequence similarity. *RA3* TPP activity was tested by a phosphate release assay with the use of recombinant *RA3* protein²³. *RA3* catalysed phosphate release from T6P but not from other sugar phosphates or a general substrate of protein phosphatases, indicating that it might act specifically as a TPP (Supplementary Fig. 4a). TPP activity was also confirmed by complementation of a yeast TPP mutant²⁴ by *RA3* (Supplementary Fig. 4b). Our data therefore indicate that *RA3* acts specifically as a TPP *in vivo*.

A *RAMOSA* pathway in maize was recently proposed in which *RAMOSA2* (*RA2*) acts upstream of *RA1* (ref. 5). To determine

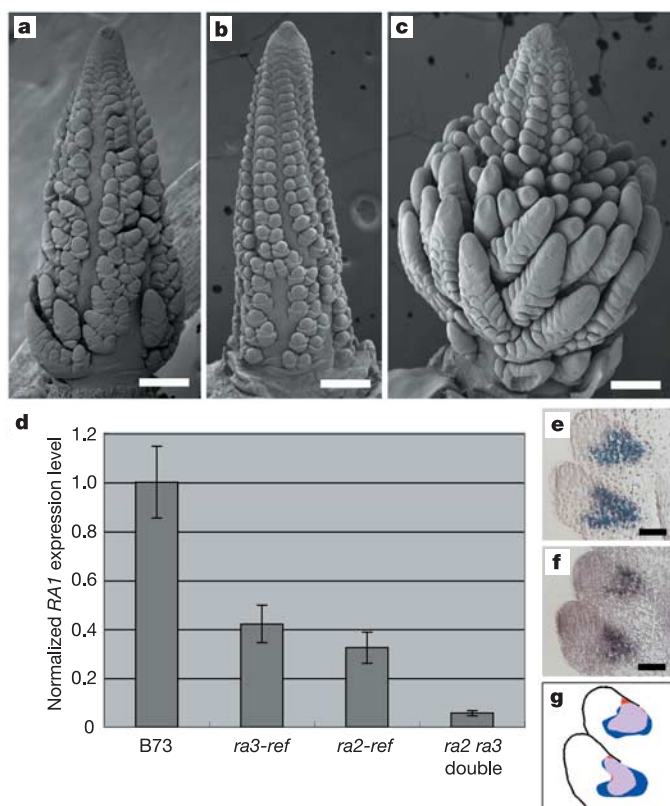


Figure 3 | *ra3* enhances a weak *ra1* mutant, and *RA1* expression is reduced in *ra3* mutants. **a**, An *ra3* mutant ear had abnormally long branches restricted to the base of the ear. **b**, The weak *ra1-RS* allele had ears with no abnormal long branches, but occasional disruption of spikelet rows. Red arrowheads show the irregular rows. **c**, A *ra3; ra1-RS* double mutant showed a strongly enhanced phenotype, with all axillary meristems growing out as long branches. Scale bars, 500 μ m. **d**, *RA1* expression was reduced in *ra2*, *ra3* and in *ra2 ra3* double mutants. Each data point is an average of three independent RT-PCR assays, and error bars indicate s.e.m. **e–g**, *RA3* (**e**) and *RA1* (**f**) are expressed in highly overlapping domains (**g**; *RA3* in blue, *RA1* in red, overlapping domain in pink). Scale bars, 50 μ m.

whether *RA3* also acts in this genetic pathway, we first examined *RA3* expression in *ra1* and *ra2* mutants. No significant change in the level or localization of *RA3* expression in these mutants was observed (data not shown). We next made double mutants between *ra3-ref* (Fig. 3a) and *ra1-RS*, a weak allele of *ra1* (ref. 5) (Fig. 3b). In ears (Fig. 3c) and tassels (not shown), the double mutants showed a strongly enhanced branching phenotype that resembled a strong *ra1* allele⁵. Expression levels of *RA1* in *ra2* mutants, in *ra3* mutants and in *ra2 ra3* double mutants in the ears were also studied. As previously reported, *RA1* expression was lower in *ra2* mutants⁵. *RA1* expression was also lower in *ra3* mutants and even lower in *ra2 ra3* double mutants (Fig. 3d), which had a similar phenotype to a strong *ra1* allele (not shown). Our data indicate that *RA3*, like *RA2*, acts upstream of *RA1* to regulate meristem identity and determinacy in maize inflorescences. Consistent with this hypothesis is the observation that *RA3* and *RA1* are expressed in overlapping domains in the developing ear (Fig. 3e–g). The sorghum *RA1* (ref. 5) and *RA3* genes are expressed similarly to maize *RA1* (ref. 5) and *RA3*, suggesting at least partial conservation of the *RAMOSA* pathway in the grasses. However, if the rice *SRA* gene also regulates inflorescence development, it must act through genes other than *RA1*, which is absent from this species⁵.

Our results indicate that trehalose metabolism can regulate a specific developmental pathway. The low endogenous levels of trehalose in most plants and the existence of multiple copies of trehalose biosynthetic genes has led to speculation that there are

spatially restricted regulatory roles for these genes⁹. Our data are also consistent with observations that the modulation of endogenous and exogenous trehalose levels affects plant growth^{13,14}. However, trehalose may act as a signal in a specific developmental pathway, because we showed that a functional TPP enzyme acts upstream of the *RA1* transcription factor to regulate inflorescence branching. The effect of *RA3* on inflorescence architecture could be mediated directly through the modulation of trehalose or T6P levels. Currently the only known targets of these sugars in plants function in metabolic signalling, in which T6P seems to be the important signal^{11,13,25,26}. The highly localized expression pattern of *RA3* and the presence of multiple TPP genes in plants mean that it may be impossible to measure differences in trehalose and T6P levels in *ra3* mutants accurately, because it is not feasible to measure these metabolites *in situ*. Indeed, our efforts so far have not succeeded in detecting reproducible differences in levels of these sugars in whole inflorescence extracts of *ra3* mutants (N.S.-N., D.J. and M. Paul, unpublished observations).

RA3 is expressed in localized domains of cells subtending, but not within, the axillary meristems in the inflorescence, indicating that it acts non-cell-autonomously. We speculate that if *RA3* does indeed act through the modulation of trehalose or T6P levels, these sugars may act as a mobile short-range signal, to regulate meristem identity and determinacy. Alternatively, our demonstration that *RA3* acts upstream of the *RA1* transcription factor would be consistent with a role for *RA3* itself in transcriptional regulation, similar to that described for some glycolytic enzymes (reviewed in ref. 27). It will be interesting to investigate further the molecular mechanism of *RA3* function, and the role of other *RA3*-like genes, to determine whether they contributed to the evolution of the unique maize inflorescence architecture that makes it one of our most successful crops^{2,5}.

METHODS

Plant materials. Plants were grown in the greenhouse or in the field, under standard conditions. The wild-type line used was B73. Inflorescences were dissected and fixed for *in situ* hybridization or imaged by SEM as described²⁸.

Mapping. Both the original *ra3-ref* allele and the *ra3-fea1* allele that we isolated from a *Mutator* transposon line were used to construct mapping populations with the wild-type line B73. Approximately 1,000 mutant individuals were used for fine mapping with molecular markers from BAC end sequences, overgoes (og; Supplementary Fig. 1) and non-repetitive sequences ('cold bands' (cb); Supplementary Fig. 1). 'Cold bands' were identified as DNA fragments on Southern blots of BAC clones that hybridized strongly to the same BAC probe but not to maize genomic DNA.

Molecular biology. Standard protocols were used for maize DNA isolation, Southern blotting and *in situ* hybridization²⁸. For labelling of *RA3* and *RA1* (ref. 5) on adjacent sections, both probes were hybridized and detected with the standard substrate²⁸ and the images were scanned into Adobe Photoshop before false colouring and superimposition. For analysis by polymerase chain reaction with reverse transcription (RT-PCR) in Supplementary Fig. 1c and Fig. 2a, poly(A)⁺ RNA was isolated with an Oligotex mRNA mini kit (Qiagen) in accordance with the manufacturer's protocol. The primers NS432 (5'-TCGTGAC AAGCATCGCAAGCA-3') and NS411 (5'-GCGCATCAGCTAGGTGTGT-3'), og15UTR (5'-ATCCATTTCATCCGTGTGGTGT-3') and og13UTR (5'-CCTGC TGACTGGACATGACTA-3') were used to amplify *RA3* and *SRA* transcripts, respectively, with the one-step RT-PCR kit (Qiagen). The control primers Ubi 5' (5'-TAAGCTGCCGATGTGCCTGCGTCG-3') and Ubi 3' (5'-CTGAAAGAC AGAACATAATGAGCAG-3') were used to amplify *UBIQUITIN* (*UBI*) transcripts. PCR conditions were 94°C for 30 s, 55°C for 30 s and 72°C for 90 s (20 cycles). The PCR cycle number was limited to ensure semiquantitative amplification, and no PCR product was visible on ethidium-bromide-stained agarose gels. The gels were Southern blotted and probed with a *RA3*, *SRA* or *UBI* probe. For RT-PCR analysis in Fig. 3d, total RNA was isolated from ears that had initiated only SPMs and semiquantitative RT-PCR was performed as described⁵. The relative *RA1* expression levels measured by RT-PCR were normalized against a ubiquitin control.

Received 16 February; accepted 15 March 2006.

1. Ward, S. P. & Leyser, O. Shoot branching. *Curr. Opin. Plant Biol.* 7, 73–78 (2004).

2. Doebley, J. Mapping the genes that made maize. *Trends Genet.* **8**, 302–307 (1992).
3. Bommert, P., Satoh-Nagasawa, N., Jackson, D. & Hirano, H. Genetics and evolution of inflorescence and flower development in grasses. *Plant Cell Physiol.* **46**, 69–78 (2005).
4. Schmitz, G. & Theres, K. Shoot and inflorescence branching. *Curr. Opin. Plant Biol.* **8**, 506–511 (2005).
5. Vollbrecht, E., Springer, P. S., Goh, L., Buckler, E. S. & Martienssen, R. Architecture of floral branch systems in maize and related grasses. *Nature* **436**, 1119–1126 (2005).
6. Strom, A. R. & Kaasen, I. Trehalose metabolism in *Escherichia coli*: stress protection and stress regulation of gene expression. *Mol. Microbiol.* **8**, 205–210 (1993).
7. Thevelein, J. M. & Hohmann, S. Trehalose synthase: guard to the gate of glycolysis in yeast? *Trends Biochem. Sci.* **20**, 3–10 (1995).
8. Goddijn, O. J. & van Dun, K. Trehalose metabolism in plants. *Trends Plant Sci.* **4**, 315–319 (1999).
9. Leyman, B., Van Dijck, P. & Thevelein, J. M. An unexpected plethora of trehalose biosynthesis genes in *Arabidopsis thaliana*. *Trends Plant Sci.* **6**, 510–513 (2001).
10. Eastmond, P. J., Li, Y. & Graham, I. A. Is trehalose-6-phosphate a regulator of sugar metabolism in plants? *J. Exp. Bot.* **54**, 533–537 (2003).
11. Schlupmann, H., Pellny, T., van Dijken, A., Smeekens, S. & Paul, M. Trehalose 6-phosphate is indispensable for carbohydrate utilization and growth in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **100**, 6849–6854 (2003).
12. Cabib, E. & Leloir, L. F. The biosynthesis of trehalose phosphate. *J. Biol. Chem.* **231**, 259–275 (1958).
13. Schlupmann, H. *et al.* Trehalose mediated growth inhibition of *Arabidopsis* seedlings is due to trehalose-6-phosphate accumulation. *Plant Physiol.* **135**, 879–890 (2004).
14. van Dijken, A. J., Schlupmann, H. & Smeekens, S. C. *Arabidopsis trehalose-6-phosphate synthase 1* is essential for normal vegetative growth and transition to flowering. *Plant Physiol.* **135**, 969–977 (2004).
15. Perry, H. S. Maize Genetics/Genomics Database [online] (<http://www.maizegdb.org/>).
16. Michmore, R. W., Paran, I. & Kesseli, R. V. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl Acad. Sci. USA* **88**, 9828–9832 (1991).
17. Thaller, M. C., Schippa, S. & Rossolini, G. M. Conserved sequence motifs among bacterial, eukaryotic, and archaeal phosphatases that define a new phosphohydrolase superfamily. *Protein Sci.* **7**, 1647–1652 (1998).
18. Plant Genome DataBase. [online] (<http://www.plantgdb.org/prj/GSSAssembly/>).
19. The Institute for Genomic Research. The TIGR Maize Database [online] (<http://www.tigr.org/tdb/tgi/maize/>).
20. Emrich, S. J. *et al.* A strategy for assembling the maize (*Zea mays* L.) genome. *Bioinformatics* **20**, 140–147 (2004).
21. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
22. Vogel, G., Aeschbacher, R. A., Muller, J., Boller, T. & Wiemken, A. Trehalose-6-phosphate phosphatases from *Arabidopsis thaliana*: identification by functional complementation of the yeast *tps2* mutant. *Plant J.* **13**, 673–683 (1998).
23. Klutts, S. *et al.* Purification, cloning, expression, and properties of mycobacterial trehalose-phosphate phosphatase. *J. Biol. Chem.* **278**, 2093–2100 (2003).
24. De Virgilio, C. *et al.* Disruption of *TPS2*, the gene encoding the 100-kDa subunit of the trehalose-6-phosphate synthase/phosphatase complex in *Saccharomyces cerevisiae*, causes accumulation of trehalose-6-phosphate and loss of trehalose-6-phosphate phosphatase activity. *Eur. J. Biochem.* **212**, 315–323 (1993).
25. Pellny, T. K. *et al.* Genetic modification of photosynthesis with *E. coli* genes for trehalose synthesis. *Plant Biotechnol. J.* **2**, 71–82 (2004).
26. Kolbe, A. *et al.* Trehalose-6-phosphate regulates starch synthesis via posttranscriptional redox activation of ADP-glucose pyrophosphorylase. *Proc. Natl Acad. Sci. USA* **102**, 11118–11123 (2005).
27. Kim, J.-w. & Dang, C. V. Multifaceted roles of glycolytic enzymes. *Trends Biochem. Sci.* **30**, 142–150 (2005).
28. Taguchi-Shiobara, F., Yuan, Z., Hake, S. & Jackson, D. The *fasciated ear2* gene encodes a leucine rich repeat receptor like protein that regulates shoot meristem proliferation in maize. *Genes Dev.* **15**, 2755–2766 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Carson and E. Coe for initial molecular mapping of *RA3*; S. Hake for the *fea1-Mu* line; N. Inada, E. Irish, J. Linder and E. Vollbrecht for *ra3* alleles; T. Mulligan for plant care; J. Andersen, N. Kobayashi-Simorowski and N. Tonks for help with phosphatase assays; V. Korothe Edavana for suggestions about the *Mycobacterium* TPP clone; P. Dahl, D. Goto, K. Noma and T. Phelps-Durr for suggestions for the yeast complementation test; J. Kossuth for help with DNA sequencing; and E. Kellogg, W. Lukowitz, J. Simorowski, E. Vollbrecht, and members of the Jackson laboratory for comments on the manuscript. Funding was provided by the National Science Foundation, Plant Genome Research Program, and the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service (to D.J.).

Author Contributions N.S.-N. performed the SEM analyses, *RA3* mapping, RT-PCRs, *in situ* hybridizations, double-mutant analyses, phosphatase assay and yeast complementation test. N.N. helped with *RA3* mapping and provided the material for RT-PCR in rice. S.M. performed phylogenetic analyses and *in situ* hybridizations in rice. H.S. organized the collaboration. D.J. supervised the research and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Accession numbers for gene sequences are listed in Supplementary Fig. 2. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.J. (jacksond@cshl.edu).

Transforming growth factor- β induces development of the T_H17 lineage

Paul R. Mangan^{1,2}, Laurie E. Harrington¹, Darrell B. O'Quinn¹, Whitney S. Helms¹, Daniel C. Bullard³, Charles O. Elson⁴, Robin D. Hatton¹, Sharon M. Wahl⁵, Trenton R. Schoeb³ & Casey T. Weaver^{1,2}

A new lineage of effector CD4⁺ T cells characterized by production of interleukin (IL)-17, the T-helper-17 (T_H17) lineage, was recently described based on developmental and functional features distinct from those of classical T_H1 and T_H2 lineages^{1,2}. Like T_H1 and T_H2, T_H17 cells almost certainly evolved to provide adaptive immunity tailored to specific classes of pathogens³, such as extracellular bacteria⁴. Aberrant T_H17 responses have been implicated in a growing list of autoimmune disorders^{5–7}. T_H17 development has been linked to IL-23, an IL-12 cytokine family member that shares with IL-12 a common subunit, IL-12p40 (ref. 8). The IL-23 and IL-12 receptors also share a subunit, IL-12R β 1, that pairs with unique, inducible components, IL-23R and IL-12R β 2, to confer receptor responsiveness⁹. Here we identify transforming growth factor- β (TGF- β) as a cytokine critical for commitment to T_H17 development. TGF- β acts to upregulate IL-23R expression, thereby conferring responsiveness to IL-23. Although dispensable for the development of IL-17-producing T cells *in vitro* and *in vivo*, IL-23 is required for host protection against a bacterial pathogen, *Citrobacter rodentium*. The action of TGF- β on naive T cells is antagonized by interferon- γ and IL-4, thus providing a mechanism for divergence of the T_H1, T_H2 and T_H17 lineages.

Interferon- γ (IFN- γ) potently inhibits T_H17 development^{1,2}. Given the suppressive actions of TGF- β on IFN- γ production^{10–12}, we speculated that TGF- β might contribute to T_H17 development by limiting inhibitory actions of IFN- γ . Naive CD4⁺ T cells were therefore activated under T_H2-neutralizing conditions and controlled availability of IL-23 and IFN- γ , with or without exogenous TGF- β 1, and cytokine phenotypes were examined (Fig. 1). Addition of IL-23 did not substantially enhance development of IL-17⁺ cells unless endogenous IFN- γ was neutralized (Fig. 1a, top panel). Addition of TGF- β 1 alone reduced the fraction of IFN- γ ⁺ T cells by more than twofold and induced the development of a small, but appreciable, fraction of IL-17⁺ cells (Fig. 1a, bottom panel). In the presence of exogenous IL-23, TGF- β 1 suppressed IFN- γ induction similarly to that in the absence of added IL-23, while modestly increasing IL-17⁺ cells. Under conditions of IL-23 addition and IFN- γ neutralization, exogenous TGF- β 1 induced further suppression of IFN- γ ⁺ cells compared to that of IL-23 addition alone and, notably, induced a markedly increased fraction of IL-17⁺ cells. Importantly, a similar induction of IL-17⁺ cells was found irrespective of exogenous IL-23 addition, suggesting that endogenous levels of IL-23 were either adequate or non-contributory. Collectively, these data indicate that in addition to its inhibitory effect on T_H1 development, TGF- β 1 promotes development of T_H17 cells.

To determine whether the augmenting effects of TGF- β 1 were due to the enhanced suppression of IFN- γ or also to IFN- γ -independent mechanisms, we examined the effects of exogenous TGF- β 1 on T_H17

development under IFN- γ -null conditions. Naive *Ifng*^{-/-} T cells were activated by *Ifng*^{-/-} splenocytes as before under T_H2-neutralizing conditions (Fig. 1b). Addition of IL-23 alone induced few IL-17⁺ cells, indicating that in the absence of IFN- γ , IL-23 alone is insufficient to promote robust T_H17 development. Addition of TGF- β 1 promoted a substantially increased fraction of IL-17⁺ cells, which was only modestly augmented by co-addition of exogenous IL-23. Therefore, TGF- β 1 can act independently of IFN- γ blockade to promote T_H17 development. Notably, reconstitution of IFN- γ -deficient cultures with high levels of exogenous IFN- γ strongly inhibited T_H17 development despite abundant TGF- β 1 and IL-23. Similar results were found for IL-4 (Supplementary Fig. S1). Thus, TGF- β , IFN- γ and IL-4 act antagonistically to specify T_H17, T_H1 or T_H2 development, respectively.

Although the foregoing experiments identified an essential role for TGF- β 1 in T_H17 development, they did not exclude the possibility that TGF- β 1 acts together with endogenous IL-23. We therefore used splenocytes from IL-12p40-deficient (*Il12b*^{-/-}; hereafter called *p40*^{-/-}) mice as a source of IL-23 (and IL-12)-deficient antigen-presenting cells (APCs) with which to examine T_H17 development under defined conditions of IL-23 availability. Without IL-23 and exogenous TGF- β 1, few IL-17-producing T cells were generated, and IL-23 alone did not restore T_H17 development (Fig. 1c). Surprisingly, addition of TGF- β 1 was sufficient to induce robust T_H17 development in the absence of IL-23, and development of IL-17-producing T cells was only modestly enhanced by co-addition of IL-23. Under more stringent conditions of IFN- γ signalling deficiency, in which IFN- γ receptor-1-deficient (*Ifngr*^{-/-}) T cells were used, more striking TGF- β 1-dependent, IL-23-independent T_H17 development was observed (Supplementary Fig. S2). Thus, TGF- β 1 acts independently of IL-23 to induce T_H17 lineage commitment.

The IL-23 receptor is a heterodimer of IL-12R β 1, which is constitutively expressed by naive T cells, and IL-23R, which is not⁹. In view of the foregoing results, we speculated that TGF- β might act proximally in T_H17 development to induce IL-23R upregulation, analogous to the induction of IL-12R β 2 by IFN- γ during T_H1 development¹³. We therefore compared expression of IL-12 and IL-23 receptors during T_H1 or T_H17 development. Naive T cells from *Ifng*^{-/-} mice were activated under neutral cytokine conditions, or under T_H17- or T_H1-polarizing conditions, and relative expression of IL-12R β 1, IL-12R β 2 and IL-23R messenger RNA determined (Fig. 1d). Compared to neutral conditions, T_H1 polarization induced IL-12R β 2 expression while inhibiting IL-23R expression. In contrast, addition of TGF- β 1 induced IL-23R expression, irrespective of IL-23 addition. Thus, TGF- β 1 and IFN- γ differentially induce mRNA for IL-23 and IL-12 receptors, respectively. TGF- β 1 thereby acts proximally in T_H17 development to confer IL-23 responsiveness.

¹Departments of ¹Pathology, ²Microbiology, ³Genomics and ⁴Medicine, University of Alabama at Birmingham, Birmingham, Alabama 35294-2170, USA. ⁵Cellular Immunology Section, Oral Infection and Immunity Branch, National Institute of Dental and Craniofacial Disease, National Institutes of Health, Bethesda, Maryland 20892-4352, USA.

The foregoing results indicated a central role for TGF- β in the initiation of T_H17 differentiation and placed TGF- β signalling proximal to IL-23 receptor expression and signalling in the T_H17 developmental programme. To confirm and extend these findings *in vivo*, we used a natural rodent pathogen, *Citrobacter rodentium*¹⁴,

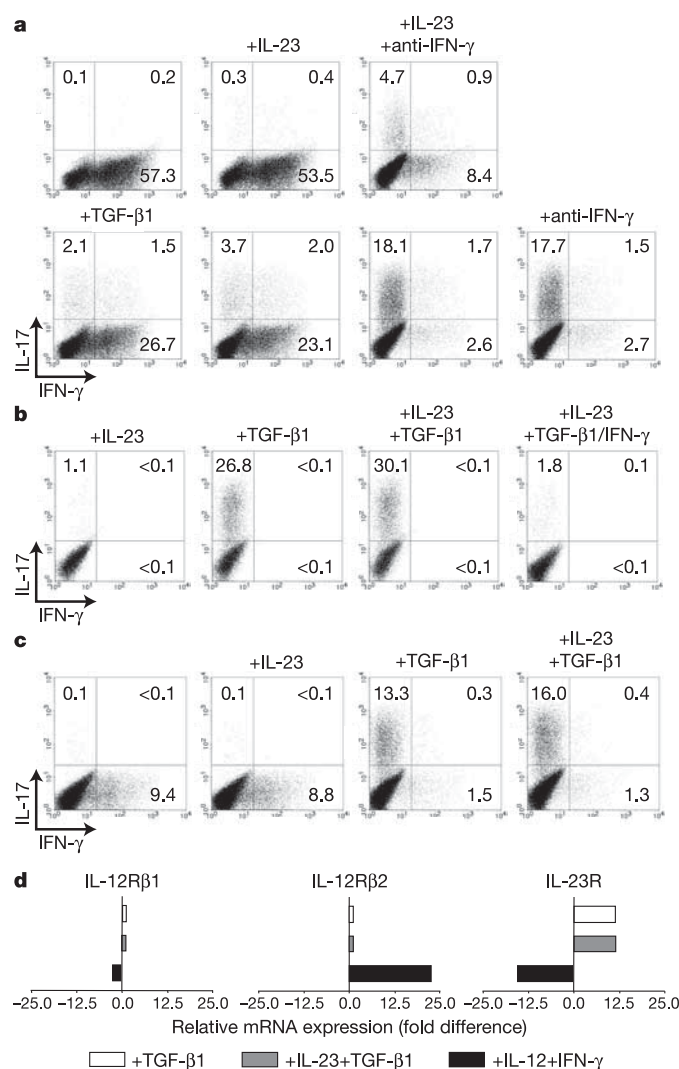


Figure 1 | TGF- β 1 is required for T_H17 commitment, independently of IL-23. **a**, Naive CD4⁺ T cells were isolated from DO11.10 TCR transgenic mice and activated with OVAp under the indicated conditions (see Supplementary Information for details). **b**, Naive CD4⁺ T cells from *Ifng*^{-/-} mice were cultured with irradiated *Ifng*^{-/-} splenocytes, anti-CD3 monoclonal antibody and anti-IL-4, with IL-23 and/or TGF- β 1, as indicated. IFN- γ was included in cultures where indicated. **c**, Naive CD4⁺ T cells were isolated from wild-type (B6) mice and cultured with *p40*^{-/-} splenocytes and anti-CD3, anti-IL-4 and anti-IFN- γ . Cultures were supplemented with nothing or IL-23 and TGF- β 1 added alone or in combination. In all cases (**a–c**), T cells were recovered after 6 days and re-stimulated with PMA plus ionomycin for 5 h with monensin block before intracellular cytokine staining for IL-17 and IFN- γ , and analysis by flow cytometry. Plots are gated on CD4⁺ cells and the quadrant percentages of cells staining positively for the indicated cytokines are shown. **d**, TGF- β and IFN- γ reciprocally induce expression of the IL-23 and IL-12 receptors. Naive CD4⁺ T cells isolated from *Ifng*^{-/-} mice were activated with *Ifng*^{-/-} splenocytes and anti-CD3 plus anti-IL-4. Cultures were supplemented with TGF- β 1 alone, TGF- β 1 plus IL-23, or IFN- γ plus IL-12, as indicated. After 6 days, T cells were collected and processed for mRNA quantification by real-time RT-PCR. Data shown are fold differences relative to T cells from a culture that was differentiated under non-polarizing (neutral) conditions (that is, was not supplemented with cytokines).

for which an intact IL-23–IL-17 axis seems to be essential for host protection (Fig. 2 and D.B.O. *et al.*, manuscript in preparation). In immunocompetent mice, *C. rodentium* induces a transient, distal colitis with resolution of lesions and clearance of the bacteria after 14–21 days, after induction of a systemic, CD4⁺ T-cell-dependent IgG response^{15–17}. Although previously associated with T_H1 adaptive immunity¹⁸, oral challenge with this organism induced a potent T_H17 response that was associated with host protection. At the peak of inflammation, 8 days after inoculation (Fig. 2a), a large fraction of colonic CD4⁺ T cells expressed IL-17, compared with a lower frequency of cells that expressed IFN- γ (Fig. 2b). Mice deficient in IL-23 (*p19*^{-/-}) failed to clear the infection, and uniformly succumbed at a rate comparable to mice deficient for both IL-12 and IL-23 (*p40*^{-/-}) (Fig. 2c). Although uninfected wild-type and IL-23-deficient mice had similar baseline histological features, the latter developed significantly less colonic inflammation after infection (Fig. 2d), despite impaired bacterial clearance. The induction of an IL-17 response was unimpaired in all examined tissues of infected

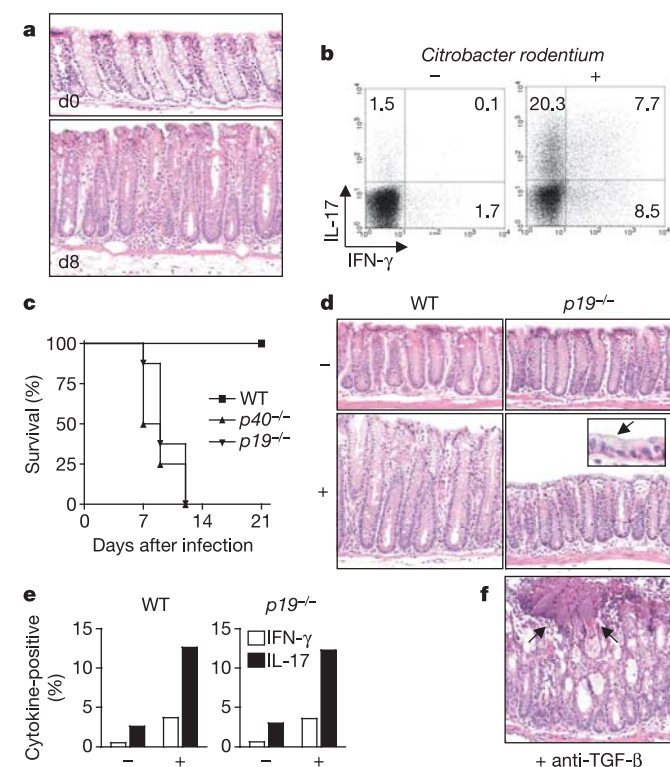


Figure 2 | In vivo development of an IL-17 effector response to a T_H17-dependent pathogen is IL-23-independent, but protection is IL-23-dependent. **a**, Histopathology of distal colon of B6 (wild-type) mice inoculated orally with $1-2 \times 10^9$ CFU *Citrobacter rodentium* and analysed before colonization (d0), or at the peak of infection (d8). **b**, Cytokine phenotype of lamina propria lymphocytes isolated from B6 mice sham-infected (-) or infected (+) with *C. rodentium* (8 days after inoculation) as in **a**. Recovered cells were processed for FACS as in Fig. 1. **c**, Survival analysis of wild-type ($n = 5$), *p40*^{-/-} ($n = 4$) and *p19*^{-/-} ($n = 8$) mice after infection with *C. rodentium* as in **a**. **d**, Histopathology of day 8 colonic tissues from sham-infected (-) or infected (+) mice of the indicated genotypes. The arrow denotes bacteria adherent to colonic epithelial cells. **e**, Intracellular cytokine analysis of lamina propria CD4⁺ T cells isolated from wild-type or *p19*^{-/-} mice 8 days after sham infection (-) or infection (+) with *C. rodentium* as in **a**. Cytokine staining and analysis as in Fig. 1. **f**, Severe ulcerative, haemorrhagic colitis induced by anti-TGF- β treatment. Histopathology of day 8 colonic tissue from *C. rodentium*-infected *p19*^{-/-} mice treated with anti-TGF- β as previously reported²⁶. The arrows denote bacteria invading ulcerated colonic epithelium. All sections (except inset) were photographed at the same magnification ($\times 20$).

IL-23-deficient mice, and dominated the IFN- γ response before and after infection (Fig. 2e, and data not shown). Thus, despite an impaired inflammatory response, and deficiencies in bacterial clearance and host protection, IL-23-deficient mice were nevertheless competent to develop a vigorous effector IL-17 response. Therefore, whereas IL-23 is dispensable for the differentiation of IL-17-competent T cells, it is indispensable for a fully effective, protective T_H17 response.

To determine whether TGF- β deficiency impaired T_H17 development *in vivo*, we initially examined the effect of anti-TGF- β neutralizing antibody on the course of *C. rodentium* infection. IL-23-deficient mice (*p19*^{-/-}) treated with anti-TGF- β developed severe ulcerative and haemorrhagic intestinal lesions with gross bacterial invasion, neither of which was found in IL-23-deficient mice treated with an isotype control, or infected wild-type mice (Fig. 2f). Although this suggested a critical role for TGF- β in protection against *C. rodentium*, a specific link with impaired T_H17 development could not be made due to the rapid morbidity and mortality associated with this treatment (data not shown). We therefore examined development of T_H17 cells in TGF- β 1-deficient mice (*Tgfb1*^{-/-}) in a non-infectious setting. In view of our finding that T_H17 cells were normally enriched in intestinal tissues (Fig. 2,

and data not shown), we surveyed T cells of the gut, as well as peripheral lymphoid tissues, for IL-17 expression (Fig. 3a). Compared to age- and sex-matched wild-type controls, we found that T_H17 cells were profoundly diminished or absent in all tissue sites of TGF- β 1-deficient mice; mice hemizygous for TGF- β 1 deficiency (*Tgfb1*^{+/-}) were intermediate. Interestingly, deficiency of IL-17⁺ cells was associated with significant decreases in basal circulating levels of IL-17 (Fig. 3b), consistent with the predominance of T cells as a source of IL-17 and a major role for the TGF- β 1 isoform in controlling homeostatic levels of IL-17 production. Furthermore, there was a striking, inverse correlation of IFN- γ -producing cells with TGF- β 1 deficiency, in accord with the spontaneous auto-inflammatory syndrome that these mice develop^{19,20}. Notably, although fewer naive precursors were present in *Tgfb1*^{-/-} mice, there was no intrinsic defect in the ability of these cells to undergo T_H17 development, provided that the high endogenous IFN- γ levels were at least partially neutralized and exogenous TGF- β 1 was provided (Supplementary Fig. S3). Collectively, these data support a critical function for TGF- β 1 in the development of T_H17 cells *in vivo*.

TGF- β 1 has been associated with immunosuppression through its inhibitory effect on effector T-cell development (for example, T_H1) and its role as an immunosuppressant cytokine produced by some regulatory T (T_{reg}) cells. TGF- β 1 also directs the development of T_{reg} cells that express the transcription factor Foxp3 (refs 21, 22). We therefore examined whether TGF- β induced expression of Foxp3 together with IL-17. Naive T cells were activated in the presence of exogenous TGF- β 1 and evaluated for intracellular expression of IL-17 and Foxp3 (Fig. 4). We found that distinct subpopulations of IL-17- and Foxp3-expressing T cells developed—albeit with a marked predominance of IL-17-producing T cells—under the conditions examined. As in previous experiments, the effects of exogenous IL-23 were modest in comparison to the effects of TGF- β 1 alone. In accordance with a recent report²³, we found that the frequency of

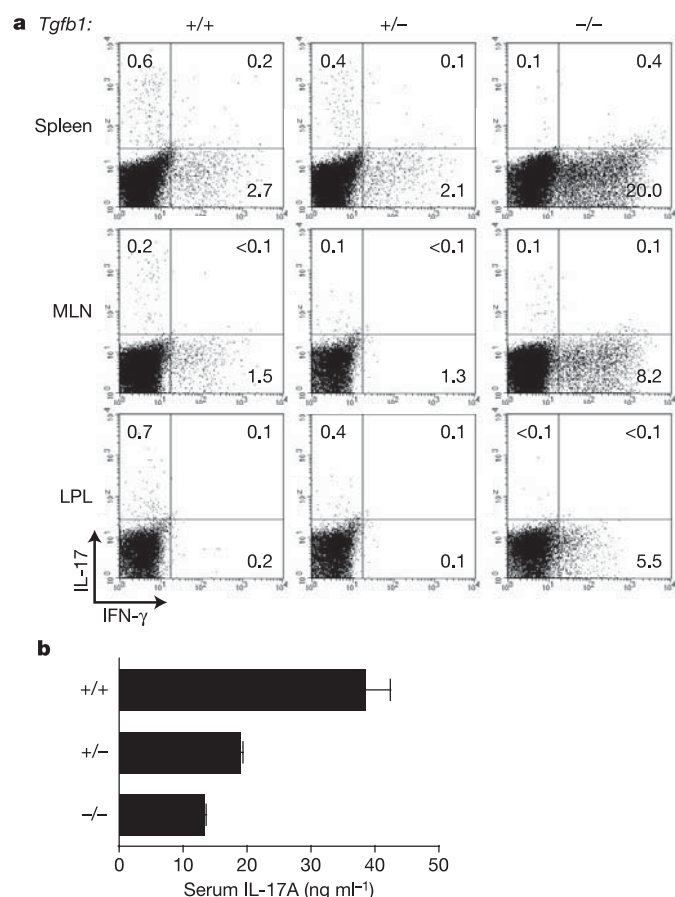


Figure 3 | Development of T_H17 cells is impaired in TGF- β 1-deficient mice. **a**, CD4⁺ T cells were purified from the indicated tissues of *Tgfb1*^{-/-} mice (-/-), and age-matched hemizygous (+/-) and wild-type (+/+) littermates. Isolated cells were stimulated immediately after isolation with PMA and ionomycin for 5 h before intracellular cytokine staining for IL-17 and IFN- γ . Stained T cells were acquired and analysed by flow cytometry, as in Fig. 1. MLN, mesenteric lymph nodes; LPL, lamina propria lymphocytes. **b**, Serum from age-matched *Tgfb1*^{-/-} mice (-/-), and hemizygous (+/-) and wild-type (+/+) littermates, was collected and analysed for IL-17A by enzyme-linked immunosorbent assay. Data are the mean $\pm$ s.e.m. of triplicate determinations from 4–8 mice.

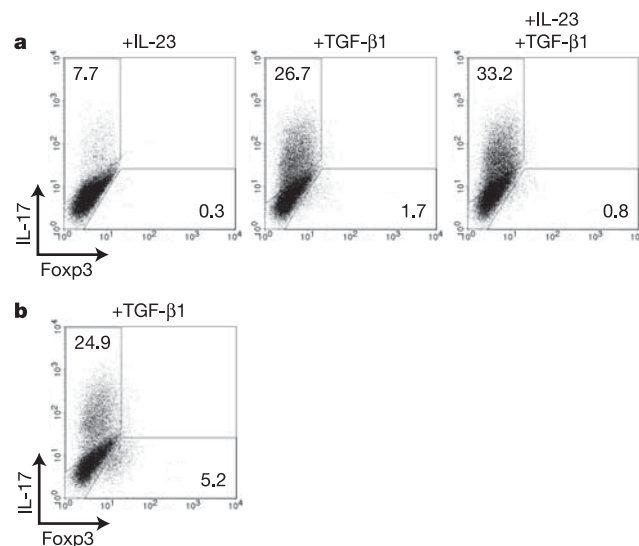


Figure 4 | TGF- β 1 induces IL-17 and Foxp3 expression by distinct CD4⁺ subpopulations. **a**, Naive CD4⁺ T cells purified from *Ifng*^{-/-} mice by FACS sorting were co-cultured with irradiated *Ifng*^{-/-} splenic feeders and activated with anti-CD3 under T_H2-neutralizing conditions. IL-23 or TGF- β 1 was added alone or in combination. **b**, Naive CD4⁺ T cells were prepared from wild-type B6 mice and cultured with wild-type splenocytes, anti-CD3, TGF- β 1, anti-IFN- γ and anti-IL-4 as in **a**. After 6 days in culture, T cells (from **a** and **b**) were re-stimulated with PMA and ionomycin before intracellular cytokine staining for IL-17 and Foxp3. Stained T cells were acquired and analysed by flow cytometry. Plots are gated on CD4⁺ cells and the quadrant percentiles are given for cells staining positively for IL-17 or Foxp3.

Foxp3⁺ cells generated in the presence of TGF- β was inversely related to levels of IL-6, such that Foxp3⁺ cells were nearly extinguished in the presence of exogenous IL-6 (Supplementary Fig. S4). Thus, TGF- β seems to have a dual role in T-cell differentiation by directing distinct subpopulations of Foxp3⁺ T_{reg} cells and T_H17 cells, contingent upon the inflammatory cytokine environment, perhaps providing a mechanism by which T_{reg} cells are poised to terminate T_H17 responses after antigen clearance.

Our findings, and complementary findings published during the revision of our paper²³, define a role for TGF- β in T_H17 lineage commitment, thereby linking this pleiotropic cytokine to adaptive immunity in a way that has important implications for mechanisms of host defence, immune homeostasis and autoimmunity. Although TGF- β has heretofore been associated primarily with immunosuppressive functions in T-cell immunity—either through the promotion of T_{reg} development and function or inhibition of T_H1 and T_H2 development¹²—it is now apparent that TGF- β may also facilitate pro-inflammatory responses by promoting T_H17 development. These data support a model in which early signalling by TGF- β in an inflammatory context initiates T_H17 commitment and upregulates IL-23R, providing a basis for TGF- β and IL-23 effects in T_H17 development that parallel those of, and are antagonized by, sequential IFN- γ and IL-12 signalling in T_H1 development. The reciprocal and antagonistic actions of TGF- β , IFN- γ and IL-4 on T_H17, T_H1 and T_H2 development—both through direct actions on the developing T cell and indirectly by modulating cytokine production by innate immune cells—provide an extended mechanism for the efficient matching of effector T-cell polarization, and thus adaptive immunity, to offending pathogens.

METHODS

CD4⁺ T-cell isolation and culture. CD4⁺ T cells from the indicated strains of mice (Supplementary Information) were purified from pooled spleen and lymph nodes by magnetic sorting using mouse anti-CD4 beads. Cells were cultured with irradiated splenic feeder cells or bone-marrow-derived dendritic cells in complete medium as described previously^{1,24}. DO11.10 T-cell receptor (TCR) transgenic CD4⁺ cells were activated with 5 μ g ml⁻¹ OVA peptide 323–339 (OVAp), whereas non-transgenic cells were stimulated with 2.5 μ g ml⁻¹ anti-CD3. Where indicated, cultures were supplemented with recombinant cytokines or antibodies (Supplementary Information). Cells were harvested on day 6 for analysis. Colonic lamina propria lymphocytes were obtained by a protocol modified from that previously described²⁵.

Citrobacter rodentium inoculation. *C. rodentium* was prepared by incubation with shaking at 37°C for 6 h in LB broth. After 6 h, the bacterial density was assessed by absorbance at an optical density of 600 nm and confirmed by plating of serial dilutions. Inoculation of mice was by oral administration with 1–2 $\times$ 10⁹ colony forming units (CFU). Tissues were collected for histology and/or cytokine phenotyping at times indicated after inoculation.

Flow cytometry. Cells were stimulated for 5–6 h with 50 ng ml⁻¹ phorbol myristate acetate (PMA) and 750 ng ml⁻¹ ionomycin, or not at all, and processed for flow cytometry as previously described¹ using the indicated antibody conjugates (see Supplementary Information). Samples were acquired on a FACSCalibur flow cytometer and data analysis used CellQuest Pro software (BD Biosciences).

RNA isolation, cDNA synthesis and real-time RT-PCR. Recovered T cells were re-stimulated with PMA and ionomycin for 5 h. Cells were lysed and RNA isolated and processed for real-time, reverse-transcribed PCR (RT-PCR) as previously described¹, using primer and probe sequences provided in Supplementary Information.

Received 6 March; accepted 28 March 2006.
Published online 30 April 2006.

- Harrington, L. E. *et al.* Interleukin 17-producing CD4⁺ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nature Immunol.* **6**, 1123–1132 (2005).
- Park, H. *et al.* A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunol.* **6**, 1133–1141 (2005).
- Murphy, K. M. & Reiner, S. L. The lineage decisions of helper T cells. *Nature Rev. Immunol.* **2**, 933–944 (2002).

- Happel, K. I. *et al.* Divergent roles of IL-23 and IL-12 in host defense against *Klebsiella pneumoniae*. *J. Exp. Med.* **202**, 761–769 (2005).
- Cua, D. J. *et al.* Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain. *Nature* **421**, 744–748 (2003).
- Murphy, C. A. *et al.* Divergent pro- and anti-inflammatory roles for IL-23 and IL-12 in joint autoimmune inflammation. *J. Exp. Med.* **198**, 1951–1957 (2003).
- Langrish, C. L. *et al.* IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. *J. Exp. Med.* **201**, 233–240 (2005).
- Oppmann, B. *et al.* Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity* **13**, 715–725 (2000).
- Parham, C. *et al.* A receptor for the heterodimeric cytokine IL-23 is composed of IL-12RB1 and a novel cytokine receptor subunit, IL-23R. *J. Immunol.* **168**, 5699–5708 (2002).
- Laouar, Y., Sutterwala, F. S., Gorelik, L. & Flavell, R. A. Transforming growth factor- β controls T helper type 1 cell development through regulation of natural killer cell interferon- γ . *Nature Immunol.* **6**, 600–607 (2005).
- Lin, J. T., Martin, S. L., Xia, L. & Gorham, J. D. TGF- β 1 uses distinct mechanisms to inhibit IFN- γ expression in CD4⁺ T cells at priming and at recall: differential involvement of Stat4 and T-bet. *J. Immunol.* **174**, 5950–5958 (2005).
- Li, M. O., Wan, Y. Y., Sanjabi, S., Robertson, A. K. & Flavell, R. A. Transforming growth factor- β regulation of immune responses. *Annu. Rev. Immunol.* **24**, 99–146 (2006).
- Berenson, L. S., Ota, N. & Murphy, K. M. Issues in T-helper 1 development—resolved and unresolved. *Immunol. Rev.* **202**, 157–174 (2004).
- MacDonald, T. T., Frankel, G., Dougan, G., Gonçalves, N. S. & Simmons, C. Host defences to *Citrobacter rodentium*. *Int. J. Med. Microbiol.* **293**, 87–93 (2003).
- Simmons, C. P. *et al.* Central role for B lymphocytes and CD4⁺ T cells in immunity to infection by the attaching and effacing pathogen *Citrobacter rodentium*. *Infect. Immun.* **71**, 5077–5086 (2003).
- Maaser, C. *et al.* Clearance of *Citrobacter rodentium* requires B cells but not secretory immunoglobulin A (IgA) or IgM antibodies. *Infect. Immun.* **72**, 3315–3324 (2004).
- Bry, L. & Brenner, M. B. Critical role of T cell-dependent serum antibody, but not the gut-associated lymphoid tissue, for surviving acute mucosal infection with *Citrobacter rodentium*, an attaching and effacing pathogen. *J. Immunol.* **172**, 433–441 (2004).
- Higgins, L. M., Frankel, G., Douce, G., Dougan, G. & MacDonald, T. T. *Citrobacter rodentium* infection in mice elicits a mucosal Th1 cytokine response and lesions similar to those in murine inflammatory bowel disease. *Infect. Immun.* **67**, 3031–3039 (1999).
- Shull, M. M. *et al.* Targeted disruption of the mouse transforming growth factor- β 1 gene results in multifocal inflammatory disease. *Nature* **359**, 693–699 (1992).
- Kulkarni, A. B. *et al.* Transforming growth factor β 1 null mutation in mice causes excessive inflammatory response and early death. *Proc. Natl Acad. Sci. USA* **90**, 770–774 (1993).
- Chen, W. *et al.* Conversion of peripheral CD4⁺CD25⁻ naive T cells to CD4⁺CD25⁺ regulatory T cells by TGF- β induction of transcription factor Foxp3. *J. Exp. Med.* **198**, 1875–1886 (2003).
- Wan, Y. Y. & Flavell, R. A. Identifying Foxp3-expressing suppressor T cells with a bicistronic reporter. *Proc. Natl Acad. Sci. USA* **102**, 5126–5131 (2005).
- Veldhoen, M., Hocking, R. J., Atkins, C. J., Locksley, R. M. & Stockinger, B. TGF- β in the context of an inflammatory cytokine milieu supports *de novo* differentiation of IL-17-producing T cells. *Immunity* **24**, 179–189 (2006).
- Kubo, T. *et al.* Regulatory T cell suppression and anergy are differentially regulated by proinflammatory cytokines produced by TLR-activated dendritic cells. *J. Immunol.* **173**, 7249–7258 (2004).
- Iqbal, N. *et al.* T helper 1 and T helper 2 cells are pathogenic in an antigen-specific model of colitis. *J. Exp. Med.* **195**, 71–84 (2002).
- Powrie, F., Carlino, J., Leach, M. W., Mauze, S. & Coffman, R. L. A critical role for transforming growth factor- β but not interleukin 4 in the suppression of T helper type 1-mediated colitis by CD45RB^{low} CD4⁺ T cells. *J. Exp. Med.* **183**, 2669–2674 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank P. Bucy, D. Chaplin, S. Schoenberger, A. Zajac and members of the Weaver laboratory for their comments and suggestions. We thank J. Oliver, M. Blake and C. Song for animal husbandry and phenotyping, and N. LeLievre for editorial assistance. We also thank R. Kastelein for provision of p19^{-/-} mice. This work was supported by grants from the NIH (to C.T.W., R.D.H. and P.R.M.), Sankyo Co. Ltd (C.T.W.) and a postdoctoral fellowship from the National Multiple Sclerosis Society (L.E.H.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.T.W. (cweaver@uab.edu).

Reciprocal developmental pathways for the generation of pathogenic effector T_H17 and regulatory T cells

Estelle Bettelli^{1*}, Yijun Carrier^{1*}, Wenda Gao^{2*}, Thomas Korn¹, Terry B. Strom², Mohamed Oukka³, Howard L. Weiner¹ & Vijay K. Kuchroo¹

On activation, T cells undergo distinct developmental pathways, attaining specialized properties and effector functions. T-helper (T_H) cells are traditionally thought to differentiate into T_H1 and T_H2 cell subsets. T_H1 cells are necessary to clear intracellular pathogens and T_H2 cells are important for clearing extracellular organisms^{1,2}. Recently, a subset of interleukin (IL)-17-producing T (T_H17) cells distinct from T_H1 or T_H2 cells has been described and shown to have a crucial role in the induction of autoimmune tissue injury^{3–5}. In contrast, CD4⁺CD25⁺Foxp3⁺ regulatory T (T_{reg}) cells inhibit autoimmunity and protect against tissue injury⁶. Transforming growth factor- β (TGF- β) is a critical differentiation factor for the generation of T_{reg} cells⁷. Here we show, using mice with a reporter introduced into the endogenous *Foxp3* locus, that IL-6, an acute phase protein induced during inflammation^{8,9}, completely inhibits the generation of Foxp3⁺ T_{reg} cells induced by TGF- β . We also demonstrate that IL-23 is not the differentiation factor for the generation of T_H17 cells. Instead, IL-6 and TGF- β together induce the differentiation of pathogenic T_H17 cells from naive T cells. Our data demonstrate a dichotomy in the generation of pathogenic (T_H17) T cells that induce autoimmunity and regulatory (Foxp3⁺) T cells that inhibit autoimmune tissue injury.

CD4⁺CD25⁺ T_{reg} cells express the forkhead/winged helix transcription factor Foxp3 (refs 10–12). Using a gene-targeting approach, we generated a mouse in which we introduced a bicistronic enhanced green fluorescent protein (EGFP) reporter into the endogenous *Foxp3* locus (Foxp3–GFP ‘knockin’ mice), allowing us to track faithfully Foxp3-expressing regulatory cells *in vivo* and study the factors that influence Foxp3 expression and T_{reg} generation. Consistent with previous observations⁷, activation of CD4⁺Foxp3[–] T cells from the Foxp3–GFP knockin mice in the presence of TGF- β induced Foxp3 expression in 50% of T cells (Fig. 1a, b). Next, we assessed the ability of pro-inflammatory cytokines to modulate Foxp3 induction via TGF- β . Notably, TGF- β -mediated conversion of CD4⁺CD25[–] cells into Foxp3⁺GFP⁺ T_{reg} cells was strongly inhibited by IL-6 (Fig. 1a, b). To understand the mechanism by which IL-6 mediated the inhibition of Foxp3 expression, we tested the cytokine production of T cells that were activated in the presence TGF- β plus IL-6, and found that these cells produced large amounts of IL-17 (Fig. 1c).

Recent studies have shown that IL-23 can differentiate T cells into T_H17 cells (refs 3–5). To address the role of IL-23 compared with TGF- β plus IL-6 in the differentiation of naive T cells into T_H17 cells, we used T cells from a myelin (myelin oligodendrocyte glycoprotein,

MOG)-specific T-cell receptor (TCR) transgenic mouse (hereafter called 2D2)¹³. *In vitro* activation of unsorted splenic T cells from the transgenic mice in the presence MOG35–55 peptide and recombinant IL-23 resulted in the generation of 6.5% of IL-17-producing CD4⁺ T cells (Fig. 2a). Because the IL-23 receptor is expressed only on activated/memory T cells^{14,15}, we wanted to investigate further whether IL-23 could guide differentiation of naive T cells into T_H17 cells. To address this question, naive 2D2 transgenic T cells (CD4⁺CD62L^{hi})

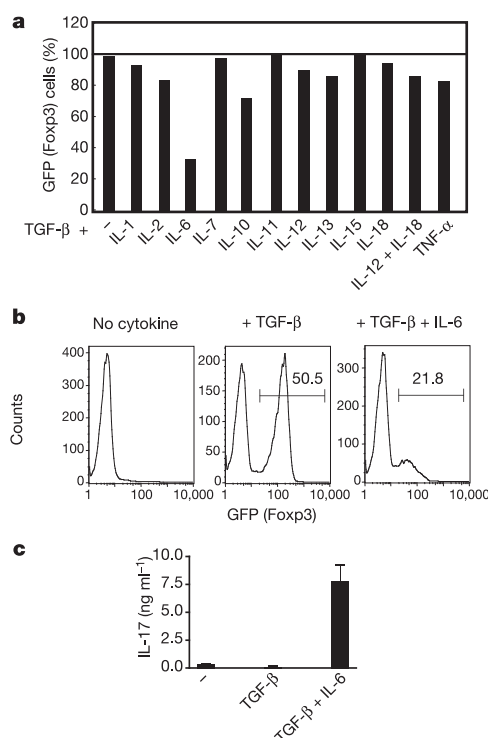


Figure 1 | Inhibition of T_{reg} development by different cytokines. **a–c**, FACS-sorted CD4⁺GFP[–] (Foxp3[–]) cells from Foxp3–GFP knockin mice were stimulated with anti-CD3 and anti-CD28 antibodies for 3 days (**a**, **b**) or anti-CD3 plus antigen-presenting cells for 48 h in the presence of different cytokines (**c**). **a**, Percentage of GFP-expressing cells induced by various cytokines in the presence of TGF- β (the percentage of GFP⁺ cells induced in the presence of TGF- β was kept at 100%). **b**, GFP expression in CD4⁺ T cells. **c**, Amount of IL-17 produced in triplicate wells $\pm$ s.d.

¹Center for Neurologic Diseases, Brigham and Women's Hospital, and ²Transplant Research Center, Beth Israel Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. ³Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, 65 Landsdowne Street, Cambridge, Massachusetts 02139, USA.

*These authors contributed equally to this work.

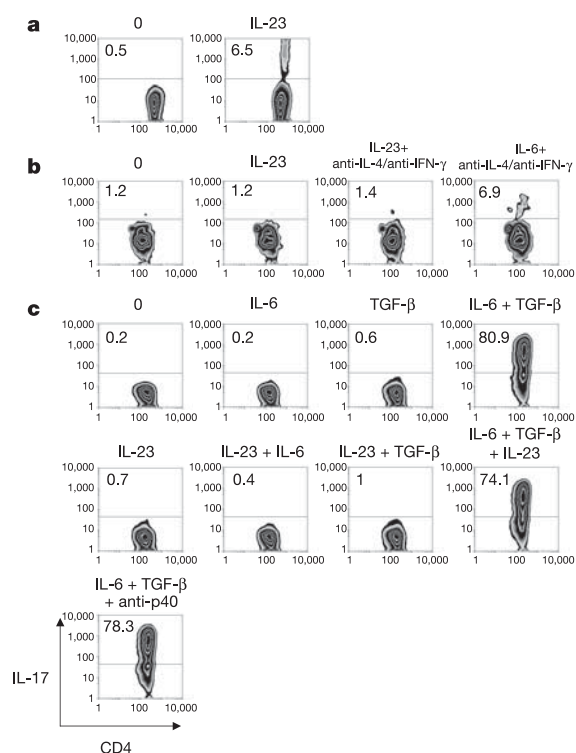


Figure 2 | Cytokines required for the generation of T_H17 cells.

a, Splenocytes from 2D2 TCR transgenic mice were stimulated with MOG35–55 peptide in neutral conditions or with IL-23. **b**, **c**, FACS-sorted naive $CD4^+CD62L^{hi}$ T cells from 2D2 TCR transgenic mice were stimulated with soluble anti-CD3, C57BL/6 irradiated spleen cells and the indicated cytokines plus neutralizing antibodies. Five days after the activation, cells were re-stimulated with PMA/ionomycin and subjected to intracellular cytokine staining. The contour plots represent IL-17 expression in $CD4^+$ cells.

were stimulated with anti-CD3 in the presence or absence of IL-23. Surprisingly, IL-23 alone, in the presence of neutralizing antibodies to IL-4 and interferon- γ (IFN- γ) (Fig. 2b), or together with other cytokines (IL-6 or TGF- β), did not induce differentiation of IL-17-producing T cells (Fig. 2c). These results suggest that IL-23 may expand the pool of IL-17-producing cells from *in vivo* activated/memory T-cell populations, but that IL-23 is incapable of differentiating IL-17-producing cells from naive T-cell precursors. Although neither TGF- β nor IL-6 alone induced the generation of IL-17-producing T cells from naive 2D2 TCR transgenic T cells, a combination of the two cytokines (IL-6 plus TGF- β), consistent with a recently published report¹⁶, induced most naive T cells (80.9%) to produce IL-17 (Fig. 2c). TGF- β could serve to inhibit IL-4 and IFN- γ production and thus allow IL-6 to induce preferentially the generation of T_H17 cells. However, culture of naive T cells with IL-6 and a cocktail of neutralizing antibodies for IL-4 and IFN- γ modestly increased (by 6.7%) the number of IL-17-producing cells compared with cells cultured with IL-6 alone, suggesting that TGF- β does not just suppress the generation of T_H1 and T_H2 cells but actively participates in the generation of T_H17 cells (Fig. 2b). In addition, IL-17 production induced by IL-6 plus TGF- β was not inhibited by the addition of a neutralizing antibody specific for the p40 chain of IL-23 (Fig. 2c, see also Supplementary Fig. S1).

We have shown that whereas IL-6 suppresses TGF- β -induced Foxp3 expression and T_{reg} generation, the combination of IL-6 and TGF- β promotes the generation of T_H17 cells, suggesting that effector and regulatory T cells may differentiate from the same precursor T cell depending on the balance of cytokines present in the environment. To address this hypothesis, we differentiated naive $CD4^+CD62L^{hi}GFP^-$ T cells from 2D2 $\times$ Foxp3–GFP knockin mice

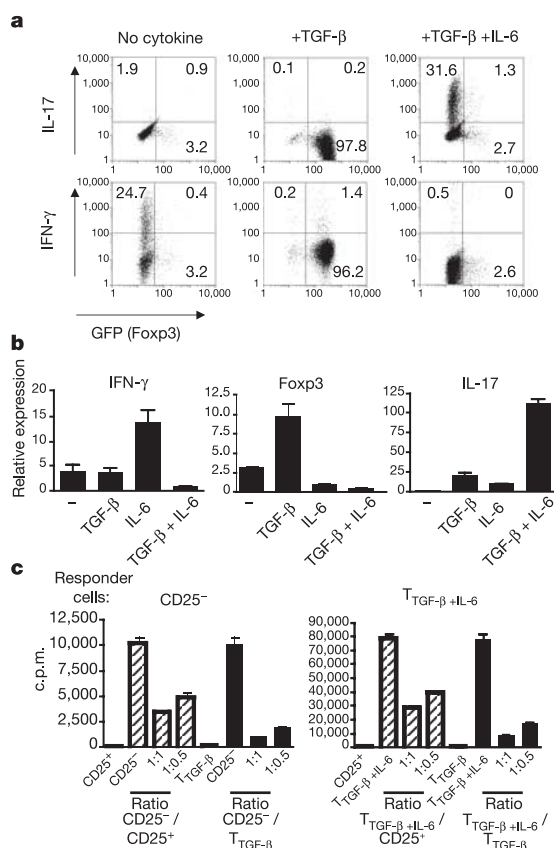


Figure 3 | Reciprocal expression of Foxp3 and IL-17 in T cells during differentiation.

FACS-sorted naive $CD4^+CD62L^{hi}GFP^-$ T cells from 2D2 $\times$ Foxp3–GFP knockin mice were stimulated with soluble anti-CD3, antigen-presenting cells and the indicated cytokines. **a**, Foxp3 (GFP) and cytokine (IL-17 or IFN- γ) intracellular expression in $CD4^+$ T cells after 5 days. **b**, Expression of IFN- γ , Foxp3 and IL-17 mRNA relative to control actin (in triplicate wells $\pm$ s.d.). **c**, Proliferation in triplicate wells ($\pm$ s.d.) of $CD4^+CD25^-$ cells and $T_{TGF-\beta+IL-6}$ cells (T cells differentiated in the presence of TGF- β and IL-6) in the presence of antigen-presenting cells, soluble anti-CD3 and different concentrations of $CD4^+CD25^-$ T cells or $T_{TGF-\beta}$ cells (T cells generated in the presence of TGF- β) measured by [³H]thymidine incorporation.

under neutral conditions, or in the presence of TGF- β or TGF- β plus IL-6, and monitored the expression of IL-17 (marker of effector T cells) versus Foxp3–GFP (marker of T_{reg} cells). Activation of T cells with anti-CD3 in the absence of exogenous cytokines did not result in the induction of either IL-17 or Foxp3, but instead resulted in the induction of some IFN- γ -producing T cells (Fig. 3a, b). Addition of TGF- β in the culture resulted in the generation of Foxp3⁺ T cells but not IL-17 or IFN- γ production; however, addition of TGF- β plus IL-6 to T cells during differentiation completely abrogated the expression of Foxp3 and resulted in concomitant expression of IL-17 from these T cells (Fig. 3a, b). No IFN- γ -producing cells were detected under these conditions. Thus, the induction of Foxp3 and IL-17 production in T cells seems to be mutually exclusive. To test the functionality of cells stimulated with TGF- β or TGF- β plus IL-6, we purified $CD4^+CD62L^{hi}GFP^-$ cells from Foxp3–GFP knockin mice and activated them in the presence of either TGF- β or TGF- β plus IL-6. Most TCR-activated $CD4^+$ T cells cultured in the presence of TGF- β became GFP positive, and the cells cultured in the presence of TGF- β plus IL-6 remained GFP negative. T cells generated in the presence of TGF- β and expressing Foxp3 (GFP⁺) were anergic and inhibited the proliferative response of $CD4^+CD25^-$ T cells similar to naturally occurring $CD4^+CD25^-$ T_{reg} cells (Fig. 3c), demonstrating that TGF- β does induce *bona fide* T_{reg} cells. In

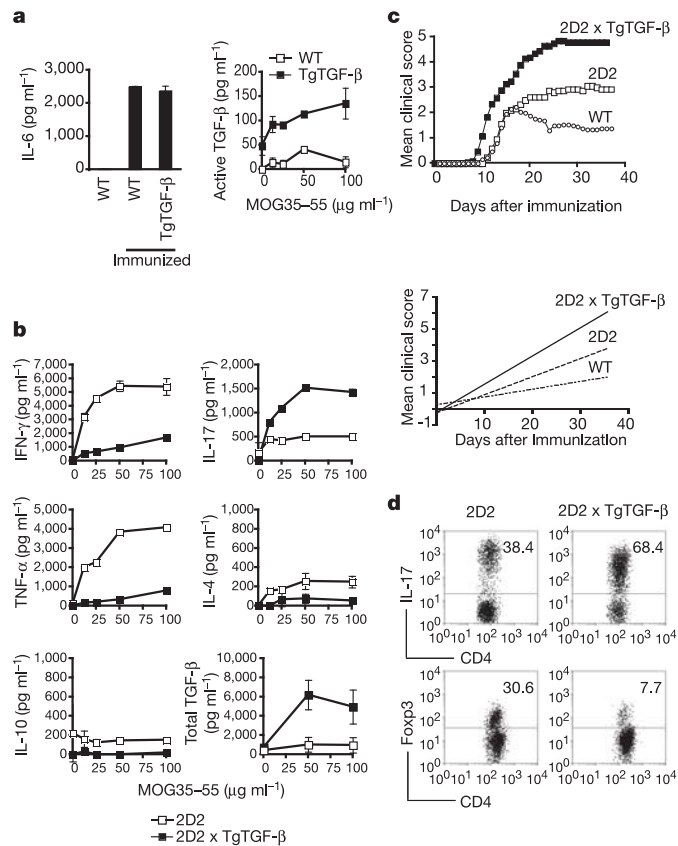


Figure 4 | Enhanced IL-17 production by T cells and CNS autoimmunity in TGF-β transgenic mice. **a**, **b**, Splenocytes from naive C57BL/6 wild-type (WT) mice (**a**, left panel), wild-type and TgTGF-β mice immunized with MOG35-55 peptide (**a**, right panel), and 2D2 and 2D2 × TgTGF-β mice immunized with MOG35-55 peptide (**b**), were stimulated with lipopolysaccharide (for IL-6) or MOG35-55 peptide (for other cytokines). Data represent mean cytokine produced in each group ($n = 3$) $\pm$ s.d. **c**, Wild-type C57BL/6 ($n = 13$), 2D2 ($n = 11$) and 2D2 × TgTGF-β ($n = 13$) mice were immunized with MOG35-55 in CFA and monitored for clinical signs of EAE (top panel). Linear regression analysis of the individual disease curves reveals a highly significant difference ($P < 0.0001$) between each of the EAE groups (bottom panel). **d**, Frequency of IL-17- and Foxp3-expressing CD4⁺ T cells isolated from the CNS of 2D2 × TgTGF-β or 2D2 mice at the peak of disease.

contrast, T cells differentiated in the presence of TGF-β and IL-6 (and that produce IL-17) were neither anergic nor immunosuppressive because they showed greater proliferation compared with the CD4⁺CD25⁻ T cells. Proliferation of T cells differentiated *in vitro* in the presence of TGF-β and IL-6 could be suppressed both by naturally occurring CD4⁺CD25⁺ T_{reg} cells and by *in vitro* TGF-β-converted Foxp3⁺ T cells (Fig. 3c).

TGF-β has been shown to be a very important immunosuppressive cytokine that not only directly suppresses effector T-cell function but

also induces CD4⁺CD25⁺ T_{reg} cells and T_H3 cells^{17,18}. To test the effect of TGF-β on T cells *in vivo*, we have generated a TGF-β transgenic mouse (TgTGF-β) in which TGF-β1 is placed under the IL-2 promoter. *In vitro* activation of 2D2 × TgTGF-β T cells with MOG35-55 peptide results in the production of large amounts of TGF-β, and these cells suppress experimental autoimmune encephalomyelitis (EAE) on adoptive transfer, showing that activation of TGF-β alone in these cells results in the generation of regulatory T cells that suppress EAE (Y.C. and H.L.W., manuscript submitted). To test the combined effect of TGF-β and IL-6 *in vivo*, we immunized TgTGF-β mice with MOG35-55 peptide emulsified in complete Freund's adjuvant (CFA). Spleen cells from immunized wild-type or TgTGF-β mice produced large amounts of IL-6 compared with spleen cells from non-immunized mice (Fig. 4a). T cells from immunized 2D2 TCR transgenic mice produced large amounts of IFN-γ, tumour-necrosis factor-α (TNF-α) and very low levels of IL-4, IL-10 and IL-17 (Fig. 4b). In contrast, T cells from MOG/CFA-immunized 2D2 × TgTGF-β mice produced TGF-β, low levels of IFN-γ and TNF-α, and very high levels of IL-17 (Fig. 4a, b). We then immunized 2D2 and 2D2 × TgTGF-β mice with MOG/CFA for the development of EAE. 2D2 TCR transgenic mice developed more severe disease compared with wild-type mice (Fig. 4c). Notably, the 2D2 × TgTGF-β mice developed the most severe and fulminant disease (Fig. 4c, Table 1 and Supplementary Table S1). Furthermore, analysis of the cytokine profile of the infiltrating CD4⁺ T cells in the central nervous system (CNS) of 2D2 versus 2D2 × TgTGF-β mice showed that the double-transgenic mice had twice as many infiltrating IL-17-producing CD4⁺ T cells in the CNS than 2D2 transgenic mice. Consistent with the idea of a reciprocal relationship between T_{reg} cells and T_H17 cells, the 2D2 × TgTGF-β mice had a reduced frequency (75% less) of Foxp3⁺ T_{reg} cells infiltrating in the CNS compared with the 2D2 transgenic mice (Fig. 4d). These *in vivo* data further show that activation of T cells in the presence of TGF-β and IL-6 results in the predominant generation of IL-17-producing T cells and fewer Foxp3⁺ T_{reg} cells and that, under these conditions, mice develop exacerbated autoimmune disease.

Our data suggest that there is not only a functional antagonism between T_H17 and T_{reg} cells but that there is a dichotomy in their generation as well. Therefore, T_{reg} cells and T_H17 effectors arise in a mutually exclusive fashion, depending on whether they are activated in the presence of TGF-β or TGF-β plus IL-6. At the steady-state level or in the absence of any inflammatory insult, TGF-β produced in the immune system will suppress the generation of effector T cells and induce Foxp3⁺ regulatory T cells, and thereby maintain self-tolerance. However, on infection or inflammation, IL-6 produced by the activated innate immune system will suppress the generation of TGF-β-induced T_{reg} cells and induce a pro-inflammatory T-cell response predominated by T_H17 cells. This is consistent with the *in vivo* observation that destructive arthritis in IFN-γ-deficient mice can be treated with anti-IL-17, and neutralization of IL-17 results in the generation of CD4⁺CD25⁺ regulatory T cells¹⁹, further supporting the concept that there is a functional antagonism between T_H17 effectors and CD4⁺CD25⁺ T_{reg} cells. Furthermore, IL-6-deficient mice have been shown to be highly resistant to the development

Table 1 | EAE in 2D2 and 2D2 × TgTGF-β mice

Group	Incidence	Mortality (%)	Mean day of onset (mean $\pm$ s.d.)	Mean maximum score (mean $\pm$ s.d.)
Wild type	12 of 13 (92.3%)	0	12.2 $\pm$ 1.99*	2.83 $\pm$ 0.33‡
2D2	8 of 11 (72.7%)	25.0	11.8 $\pm$ 1.91†	4.17 $\pm$ 0.68§
2D2 × TgTGF-β	13 of 13 (100%)	90.9	10.1 $\pm$ 1.50*†	4.91 $\pm$ 0.30‡§

Mice were immunized with MOG35-55 peptide emulsified in complete Freund's adjuvant and monitored for the development of EAE. Statistical analysis was performed by comparing three groups using one-way analysis of variance (ANOVA) followed by a Newman-Keuls multiple comparison as a post-hoc test.

* $P < 0.05$.

† $P < 0.05$.

‡ $P < 0.001$.

§ $P < 0.01$.

of EAE^{20–23}. We have confirmed this observation and further determined that IL-6-deficient mice immunized with the encephalitogenic MOG35–55 peptide have a deficit in IL-17-producing T cells infiltrating the CNS (Supplementary Fig. S2).

Our results support the concept that there is a reciprocal developmental pathway for the generation of pathogenic T_H17 cells and protective T_{reg} cells in the immune system depending on the state of the innate immune system and production of acute phase proteins like IL-6. Once induced, CD4⁺CD25⁺Foxp3⁺ cells could also suppress other T-cell subsets (like T_H1 and T_H2 cells), as well as T_H17 cells. T_H17 and T_{reg} subsets may therefore have evolved to induce or regulate tissue inflammation, analogous to the dichotomy of T_H1 and T_H2 T-cell subsets, which primarily mediate immunity against infectious organisms.

METHODS

Mice. Foxp3–GFP knockin, 2D2 MOG TCR and TGF- β transgenic mice were all generated on the C57BL/6 background as described in Supplementary Methods.

In vitro T-cell differentiation. Spleen cells from 2D2 mice were stimulated with 50 $\mu\text{g ml}^{-1}$ MOG35–55 peptide in the presence or absence of IL-23 (20 ng ml^{-1}). CD4⁺ T cells were purified using anti-CD4 beads (Myltenyi) or further sorted into naive CD4⁺CD62L^{hi} cells. CD4⁺ T cells were stimulated with C57BL/6 irradiated spleen cells and 1 $\mu\text{g ml}^{-1}$ of anti-CD3 (145-2C11) for 3–5 days in the presence of cytokines (human TGF- β 1 (3 ng ml^{-1}), mouse IL-6 (20 ng ml^{-1}), IL-23 (20 ng ml^{-1}) (all R&D Systems)) and/or neutralizing antibodies (anti-IL-4 (10 $\mu\text{g ml}^{-1}$, 11B11), anti-IFN- γ (10 $\mu\text{g ml}^{-1}$, XMGI.2), anti-IL-12/IL-23 p40 (C17.8)). Cells were supplemented with recombinant IL-2 (50 U ml^{-1}) at day 2 and 4.

Suppression assay. Proliferation was determined by [³H]thymidine incorporation as described in Supplementary Methods.

Cytokine production and intracellular cytokine staining. Cytokine production was determined by enzyme-linked immunosorbent assay as described in Supplementary Methods. For intracellular cytokine staining, T cells were restimulated, 5 days after their initial stimulation, with PMA/ionomycin (see Supplementary Methods) and stained according to the manufacturer's directions (BD).

Real-time PCR. The expression of IL-17, IFN- γ and Foxp3 was performed using specific primers and probes (Applied Biosystems). Expression was normalized to the expression of the housekeeping gene actin.

Induction and assessment of EAE. 2D2, 2D2 $\times$ TgTGF- β or C57BL/6 mice were injected subcutaneously with 100 μg of MOG35–55 peptide (MEVGWYRSPFSRVVHLYRNGK) emulsified in CFA (Difco) supplemented with 400 $\mu\text{g ml}^{-1}$ *Mycobacterium tuberculosis* and injected twice intravenously with 150 ng of pertussis toxin (List Biological Laboratories). Clinical assessment of EAE was performed daily after disease induction according to the following criteria: 0, no disease; 1, decreased tail tone; 2, hindlimb weakness or partial paralysis; 3, complete hindlimb paralysis; 4, forelimb and hindlimb paralysis; 5, moribund state.

Received 2 February; accepted 30 March 2006.

Published online 30 April 2006.

1. Mosmann, T. R. & Coffman, R. L. T_H1 and T_H2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* **7**, 145–173 (1989).
2. Bottomly, K. A functional dichotomy in CD4⁺ T lymphocytes. *Immunol. Today* **9**, 268–274 (1988).
3. Langrish, C. L. et al. IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. *J. Exp. Med.* **201**, 233–240 (2005).
4. Harrington, L. E. et al. Interleukin 17-producing CD4⁺ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nature Immunol.* **6**, 1123–1132 (2005).

5. Park, H. et al. A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunol.* **6**, 1133–1141 (2005).
6. Sakaguchi, S. Naturally arising CD4⁺ regulatory T cells for immunologic self-tolerance and negative control of immune responses. *Annu. Rev. Immunol.* **22**, 531–562 (2004).
7. Chen, W. et al. Conversion of peripheral CD4⁺CD25⁺ naive T cells to CD4⁺CD25⁺ regulatory T cells by TGF- β induction of transcription factor Foxp3. *J. Exp. Med.* **198**, 1875–1886 (2003).
8. Hirano, T. Interleukin 6 and its receptor: ten years later. *Int. Rev. Immunol.* **16**, 249–284 (1998).
9. Ozato, K., Tsujimura, H. & Tamura, T. Toll-like receptor signaling and regulation of cytokine gene expression in the immune system. *Biotechniques* **33** (Suppl.), S66–S68 (2002).
10. Hori, S., Nomura, T. & Sakaguchi, S. Control of regulatory T cell development by the transcription factor Foxp3. *Science* **299**, 1057–1061 (2003).
11. Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of CD4⁺CD25⁺ regulatory T cells. *Nature Immunol.* **4**, 330–336 (2003).
12. Khattri, R., Cox, T., Yasayko, S. A. & Ramsdell, F. An essential role for Scurfin in CD4⁺CD25⁺ T regulatory cells. *Nature Immunol.* **4**, 337–342 (2003).
13. Bettelli, E. et al. Myelin oligodendrocyte glycoprotein-specific T cell receptor transgenic mice develop spontaneous autoimmune optic neuritis. *J. Exp. Med.* **197**, 1073–1081 (2003).
14. Aggarwal, S., Ghilardi, N., Xie, M. H., de Sauvage, F. J. & Gurney, A. L. Interleukin-23 promotes a distinct CD4 T cell activation state characterized by the production of interleukin-17. *J. Biol. Chem.* **278**, 1910–1914 (2003).
15. Parham, C. et al. A receptor for the heterodimeric cytokine IL-23 is composed of IL-12R β 1 and a novel cytokine receptor subunit, IL-23R. *J. Immunol.* **168**, 5699–5708 (2002).
16. Veldhoen, M., Hocking, R. J., Atkins, C. J., Locksley, R. M. & Stockinger, B. TGF β in the context of an inflammatory cytokine milieu supports *de novo* differentiation of IL-17-producing T cells. *Immunity* **24**, 179–189 (2006).
17. Gorelik, L. & Flavell, R. A. Transforming growth factor- β in T-cell biology. *Nature Rev. Immunol.* **2**, 46–53 (2002).
18. Shull, M. M. et al. Targeted disruption of the mouse transforming growth factor- β 1 gene results in multifocal inflammatory disease. *Nature* **359**, 693–699 (1992).
19. Nardelli, D. T. et al. Association of CD4⁺ CD25⁺ T cells with prevention of severe destructive arthritis in *Borrelia burgdorferi*-vaccinated and challenged gamma interferon-deficient mice treated with anti-interleukin-17 antibody. *Clin. Diagn. Lab. Immunol.* **11**, 1075–1084 (2004).
20. Samoilova, E. B., Horton, J. L., Hilliard, B., Liu, T. S. & Chen, Y. IL-6-deficient mice are resistant to experimental autoimmune encephalomyelitis: roles of IL-6 in the activation and differentiation of autoreactive T cells. *J. Immunol.* **161**, 6480–6486 (1998).
21. Okuda, Y. et al. IL-6 plays a crucial role in the induction phase of myelin oligodendrocyte glycoprotein 35–55 induced experimental autoimmune encephalomyelitis. *J. Neuroimmunol.* **101**, 188–196 (1999).
22. Okuda, Y., Sakoda, S., Saeki, Y., Kishimoto, T. & Yanagihara, T. Enhancement of Th2 response in IL-6-deficient mice immunized with myelin oligodendrocyte glycoprotein. *J. Neuroimmunol.* **105**, 120–123 (2000).
23. Eugster, H. P. et al. Superantigen overcomes resistance of IL-6-deficient mice towards MOG-induced EAE by a TNFR1 controlled pathway. *Eur. J. Immunol.* **31**, 2302–2312 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. A. Sobel for histological analysis of CNS tissues from mice, D. Kozoriz for cell sorting, and A. Jäger for technical assistance. This work was supported by grants from the National Multiple Sclerosis Society, the National Institutes of Health, JDRF Center for Immunological Tolerance at Harvard and the Deutsche Forschungsgemeinschaft.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.K.K. (vkuchroo@rics.bwh.harvard.edu) and M.O. (moukka@rics.bwh.harvard.edu).

Hexon-chimaeric adenovirus serotype 5 vectors circumvent pre-existing anti-vector immunity

Diane M. Roberts¹, Anjali Nanda¹, Menzo J. E. Havenga², Peter Abbink¹, Diana M. Lynch¹, Bonnie A. Ewald¹, Jinyan Liu¹, Anna R. Thorner¹, Patricia E. Swanson¹, Darci A. Gorgone¹, Michelle A. Lifton¹, Angelique A. C. Lemckert², Lennart Holterman², Bing Chen³, Athmanundh Dilraj⁴, Angela Carville⁵, Keith G. Mansfield⁵, Jaap Goudsmit² & Dan H. Barouch¹

A common viral immune evasion strategy involves mutating viral surface proteins in order to evade host neutralizing antibodies. Such immune evasion tactics have not previously been intentionally applied to the development of novel viral gene delivery vectors that overcome the critical problem of anti-vector immunity. Recombinant, replication-incompetent adenovirus serotype 5 (rAd5) vector-based vaccines for human immunodeficiency virus type 1 and other pathogens have proved highly immunogenic in preclinical studies^{1,2} but will probably be limited by the high prevalence of pre-existing anti-Ad5 immunity in human populations, particularly in the developing world^{3–6}. Here we show that rAd5 vectors can be engineered to circumvent anti-Ad5 immunity. We constructed novel chimaeric rAd5 vectors in which the seven short hypervariable regions (HVRs) on the surface of the Ad5 hexon protein were replaced with the corresponding HVRs from the rare adenovirus serotype Ad48. These HVR-chimaeric rAd5 vectors were produced at high titres and were stable through serial passages *in vitro*. HVR-chimaeric rAd5 vectors expressing simian immunodeficiency virus Gag proved comparably immunogenic to parental rAd5 vectors in naive mice and rhesus monkeys. In the presence of high levels of pre-existing anti-Ad5 immunity, the immunogenicity of HVR-chimaeric rAd5 vectors was not detectably suppressed, whereas the immunogenicity of parental rAd5 vectors was abrogated. These data demonstrate that functionally relevant Ad5-specific neutralizing antibodies are focused on epitopes located within the hexon HVRs. Moreover, these studies show that recombinant viral vectors can be engineered to circumvent pre-existing anti-vector immunity by removing key neutralizing epitopes on the surface of viral capsid proteins. Such chimaeric viral vectors may have important practical implications for vaccination and gene therapy.

Anti-vector immunity represents a key limitation of current recombinant viral gene delivery vectors. Pre-existing anti-vector immunity has been shown to suppress the immunogenicity of rAd5 vector-based vaccines, and the generation of anti-vector immunity after immune priming has been demonstrated to limit the efficiency of homologous boost immunizations^{3,7,8}. Our laboratory and others have recently reported that dominant Ad5-specific neutralizing antibodies are directed primarily against the Ad5 hexon major capsid protein^{6,9}, suggesting the potential utility of hexon modification strategies to construct modified rAd5 vectors that circumvent anti-Ad5 immunity. Hexon exchanges among adenoviruses from different virus subgroups, however, have proven limited by viral structural constraints^{9,10}. We sought to alter the antigenicity

of the Ad5 hexon without perturbing its core structure by exchanging only the seven short HVRs on the surface of the Ad5 hexon protein, as these HVRs contain the majority of sequence variability among adenovirus serotypes¹¹. The crystal structure of the hexon protein reveals a homotrimer of subunits each consisting of a double barrel core, although the HVRs are not fully resolved in the refined structure^{12,13}. We therefore modelled the HVRs with reasonable geometry with respect to the rest of the molecule in O¹⁴ and produced ribbon diagrams and space-filling models using MolScript¹⁵. This model suggests that the HVRs in fact define most of the solvent-exposed surface of the hexon trimer and thus may contain key Ad5-specific neutralizing antibody epitopes (Fig. 1a).

We constructed chimaeric, replication-incompetent, E1/E3-deleted rAd5 vectors in which either the first and most diverse hexon HVR, or alternatively all seven hexon HVRs, were specifically replaced with the corresponding HVRs from Ad48, which has particularly low seroprevalence in humans⁴ (Fig. 1b). We termed these vectors rAd5HVR48(1) and rAd5HVR48(1–7), respectively. We did not alter the sequences of the hexon regions between the HVRs, as they form key elements of secondary structure within the hexon core and thus may be critical for hexon folding and virus stability. Consistent with these observations, our attempts to date to exchange larger regions or entire domains of the Ad5 hexon have failed to rescue virus (data not shown). In contrast, rAd5HVR48(1) and rAd5HVR48(1–7) vectors grew efficiently to high titres, although final yields of these chimaeric vectors were still three- to fivefold lower than yields of the parental rAd5 vectors. Transgene expression from rAd5HVR48(1) and rAd5HVR48(1–7) vectors expressing simian immunodeficiency virus (SIV) Gag proved comparable with that of rAd5 vectors, as measured by infection of A549 cells with varying amounts of virus followed by analysis of cell lysates by enzyme-linked immunosorbent assay (ELISA) (Fig. 1c). Moreover, specific infectivities of the chimaeric vectors were similar to those of rAd5 vectors, with virus particle to plaque-forming unit ratios <30. In addition, rAd5HVR48(1–7)-Gag vectors remained stable for at least 15 serial passages *in vitro* without detectable loss of the transgene (Fig. 1d) or changes to the chimaeric hexon sequence (data not shown).

We assessed the immunogenicity of rAd5, rAd5HVR48(1) and rAd5HVR48(1–7) vectors expressing SIV Gag in C57/BL6 mice either with or without anti-Ad5 immunity. Groups of mice were pre-immunized twice with 10¹⁰ virus particles of rAd5-Empty to generate high levels of anti-Ad5 immunity. These mice had Ad5-specific neutralizing antibody titres of 8,192–16,384, which represent the

¹Division of Viral Pathogenesis, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA. ²Crucell Holland BV, 2301 CA, Leiden, The Netherlands. ³Laboratory of Molecular Medicine, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴South African Medical Research Council, Durban 4001, South Africa. ⁵New England Primate Research Center, Southborough, Massachusetts 01772, USA.

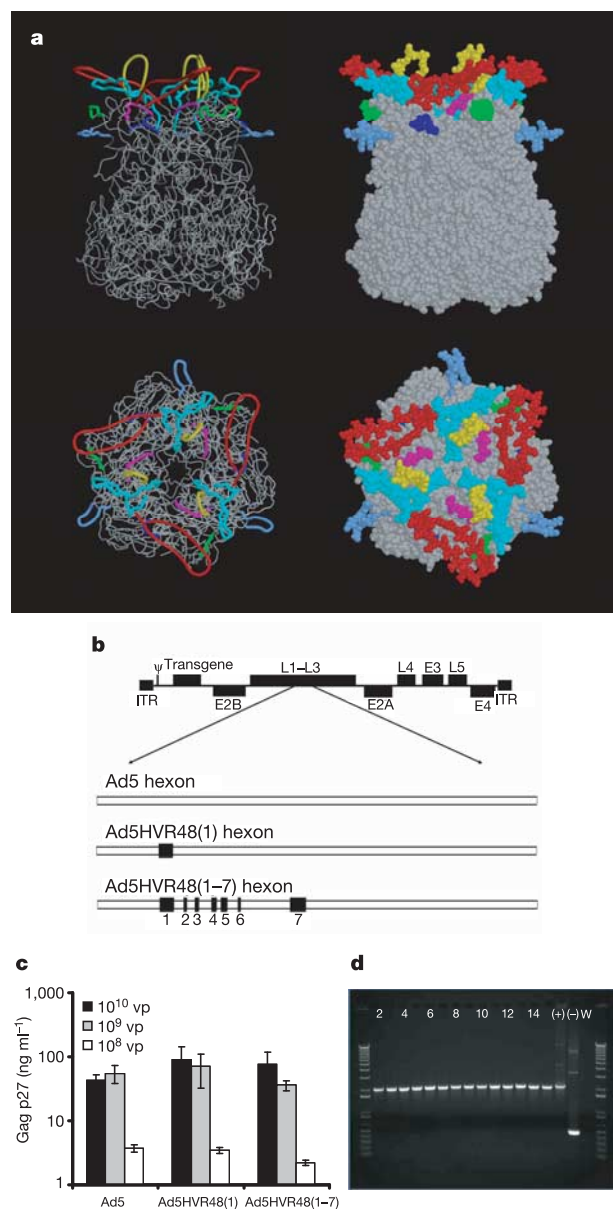


Figure 1 | Construction of hexon HVR-chimaeric rAd5 vectors. **a**, Ad5 hexon trimer structure (Protein Data Bank 1P30)¹³ with modelled and highlighted HVRs (HVR1, red; HVR2, green; HVR3, pink; HVR4, light blue; HVR5, yellow; HVR6, blue; HVR7, cyan). The remainder of the hexon trimer is shown in grey. The HVR1 loop, for which there are essentially no experimental constraints, was modelled arbitrarily, maintaining stereochemically correct geometry and avoiding self-collision. Ribbon diagrams (left panels) and space-filling models (right panels) are shown for both side views (upper panels) and top views (lower panels) of a hexon trimer. **b**, Schematic cloning strategy to construct rAd5HVR48(1) and rAd5HVR48(1-7) vectors by replacing either the first HVR or all seven HVRs, respectively, with the corresponding regions from Ad48. **c**, SIV Gag expression from rAd5-Gag, rAd5HVR48(1)-Gag and rAd5HVR48(1-7)-Gag vectors. Gag p27 ELISA analysis of cell lysates is shown after infection of A549 cells with 10^{10} , 10^9 , or 10^8 virus particles (vp) of each vector. Error bars are means $\pm$ s.e.m. **d**, Transgene stability of rAd5HVR48(1-7)-Gag vectors through 15 serial passages. PCR reactions were performed to amplify the transgene region using viral DNA extracted from crude lysates from passages 2–15 and then analysed by 0.8% agarose gel electrophoresis. PCR controls included the plasmid pAdApt-Gag (+), the plasmid pAdApt-Empty (–) and water (W).

upper bound of neutralizing antibody titres found in humans in sub-Saharan Africa^{5,6,16}. Naive mice (Fig. 2a, c, e, g) and mice with anti-Ad5 immunity (Fig. 2b, d, f, h) ($n = 4$ per group) were immunized intramuscularly with a single injection of 10^9 , 10^8 , 10^7 , or 10^6 virus particles of rAd5-Gag, rAd5HVR48(1)-Gag or rAd5HVR48(1-7)-Gag. Gag-specific CD8⁺ T lymphocyte responses specific for the dominant D^b-restricted AL11 epitope (AAVKNWMTQTLL)⁷ were assessed by D^b/AL11 tetramer binding assays. All three vectors proved comparably immunogenic in naive mice, even at the dose-limiting level of 10^7 virus particles (Fig. 2e). These data show that modifying the hexon HVRs does not diminish the potent cellular immune responses elicited by rAd5 vectors.

In mice with anti-Ad5 immunity, the immunogenicity of rAd5-Gag was abrogated at all doses tested (Fig. 2b, d, f). The immunogenicity of rAd5HVR48(1)-Gag was similarly suppressed, indicating that exchanging only the first hexon HVR was insufficient to circumvent high levels of anti-Ad5 immunity. The immunogenicity of rAd5HVR48(1-7)-Gag, however, remained essentially unaffected by anti-Ad5 immunity, even at the dose-limiting level of 10^7 virus particles (Fig. 2f). In fact, responses elicited by rAd5HVR48(1-7)-Gag were significantly higher than those elicited by rAd5-Gag or rAd5HVR48(1)-Gag in mice with anti-Ad5 immunity ($P < 0.001$, comparing tetramer binding responses on day 21 by analysis of variance (ANOVA) with Bonferroni adjustments to account for multiple comparisons). These data demonstrate that simultaneously exchanging all seven hexon HVRs results in a vector that effectively evades anti-Ad5 immunity in mice. Functional interferon (IFN)- γ ELISPOT assays in response to a pool of overlapping Gag peptides, the CD8⁺ T lymphocyte epitopes AL11 and KV9, and the CD4⁺ T lymphocyte epitope DD13 confirmed these tetramer binding assays (Fig. 2i, j).

To evaluate further the immunogenicity of the chimaeric rAd5HVR48(1-7) vector, we compared the immunogenicity of rAd5-Gag and rAd5HVR48(1-7)-Gag with the rare serotype vectors rAd35-Gag^{4,7} and rAd48-Gag. In naive mice, rAd5-Gag and rAd5HVR48(1-7)-Gag proved significantly more immunogenic than rAd35-Gag and rAd48-Gag at the lower dose of 10^7 virus particles (Fig. 2k; $P < 0.001$), consistent with the lower immunogenicity of rare serotype recombinant adenovirus vectors as compared with rAd5 vectors in preclinical studies to date^{3,7}. In mice with anti-Ad5 immunity, 10^7 virus particles of rAd5HVR48(1-7)-Gag proved significantly more immunogenic than rAd5, rAd35 and rAd48 vectors (Fig. 2l; $P < 0.001$), reflecting the intrinsic potency of this chimaeric vector combined with its capacity to circumvent anti-Ad5 immunity.

We next evaluated the immunogenicity of heterologous recombinant adenovirus prime-boost regimens in mice either with or without anti-Ad5 immunity. Groups of mice ($n = 4$ per group) were primed on day 0 with 10^9 virus particles of rAd35-Gag and then boosted on day 28 with 10^9 virus particles of rAd5HVR48(1-7)-Gag, rAd35-Gag, or rAd5-Gag. Naive mice primed with rAd35-Gag were boosted efficiently and comparably by the heterologous vectors rAd5HVR48(1-7)-Gag and rAd5-Gag (Fig. 3a, c). In contrast, re-administration of the homologous vector rAd35-Gag proved ineffective, presumably due to the generation of anti-Ad35 immunity by the priming immunization. In mice with pre-existing anti-Ad5 immunity (Fig. 3b, d), rAd5HVR48(1-7)-Gag proved significantly more potent than both rAd5-Gag and rAd35-Gag as a boosting vector ($P < 0.001$, comparing tetramer binding responses on day 49). These data suggest that rAd5HVR48(1-7) vectors function essentially as a novel serotype distinct from rAd5 vectors. Moreover, these studies show that rAd5HVR48(1-7) vectors remain highly immunogenic both as a prime and as a boost in the presence of high levels of anti-Ad5 immunity that ablate the immunogenicity of rAd5 vectors.

We measured vector-specific neutralizing antibody titres in mice immunized either once (Fig. 4a) or twice (Fig. 4b) with 10^{10} virus

particles of rAd5-Gag, rAd5HVR48(1–7)-Gag, or rAd35-Gag. As expected, mice that received rAd5-Gag developed high titres of Ad5-specific neutralizing antibodies but substantially lower titres of Ad5HVR48(1–7)-specific neutralizing antibodies, confirming that

the majority of Ad5-specific neutralizing antibodies are directed against the hexon HVRs. Conversely, mice that received rAd5HVR48(1–7)-Gag developed high titres of Ad5HVR48(1–7)-specific neutralizing antibodies but markedly lower titres of Ad5-specific neutralizing antibodies. These mice also developed high titres of Ad48-specific neutralizing antibodies, suggesting that the HVRs in this chimaeric vector are in fact presented in conformations comparable with their wild-type orientations in Ad48. Thus, exchanging the HVRs effectively swapped the predominant serologic determinants of the virus from Ad5 to Ad48. We also measured Gag-specific antibody responses in these mice by ELISA (Fig. 4c). High titres of Gag-specific antibodies were elicited by both rAd5-Gag and rAd5HVR48(1–7)-Gag, although there was a trend towards threefold lower antibody titres elicited by rAd5HVR48(1–7)-Gag. In contrast, rAd35-Gag failed to generate detectable Gag-specific antibodies in this system, consistent with our previous observations⁷.

We next assessed vector-specific neutralizing antibody titres in 265 serum samples from healthy individuals in South Africa (Fig. 4d). The median Ad5-specific neutralizing antibody titre in these samples was 1,024, whereas the median Ad5HVR48(1–7)-specific neutralizing antibody titre was significantly lower at 128 ($P < 0.0001$, Wilcoxon rank-sum test). These data suggest that 85–90% of Ad5-specific neutralizing antibodies in humans are directed against the hexon HVRs. The residual low-titre neutralizing antibodies against rAd5HVR48(1–7) presumably represent Ad5 fibre- and penton-specific neutralizing antibodies, because Ad48-specific neutralizing antibodies were particularly low in these samples. In previous adoptive transfer studies in mice, we demonstrated that Ad5 fibre- and penton-specific neutralizing antibodies did not efficiently suppress rAd5 vaccine immunogenicity⁶, consistent with the observations in the present study that rAd5HVR48(1–7) vectors effectively circumvented anti-Ad5 immunity. Moreover, serum samples with Ad48- but not Ad5-specific neutralizing antibodies also exhibited neutralizing activity against rAd5HVR48(1–7) (data not shown), supporting the conclusion that dominant neutralizing antibodies are directed primarily against the hexon HVRs.

To confirm the murine immunogenicity studies, we performed a

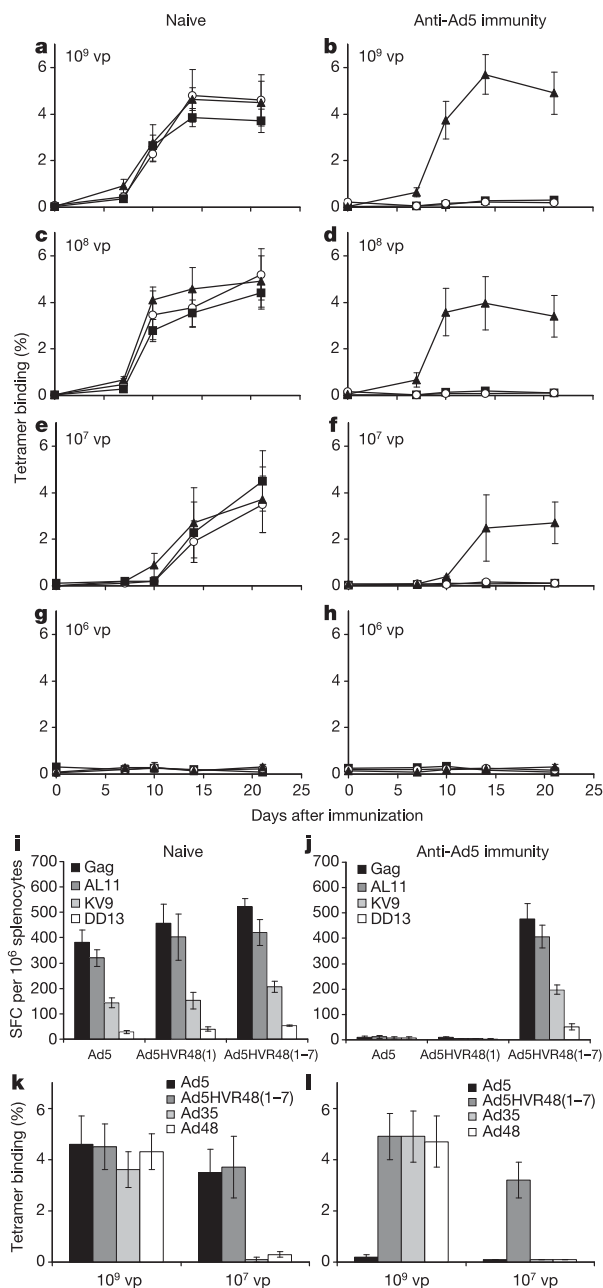


Figure 2 | Cellular immune responses elicited by hexon HVR-chimaeric rAd5 vectors. Naive C57/BL6 mice (**a, c, e, g**) or C57/BL6 mice with anti-Ad5 immunity (**b, d, f, h**) were immunized intramuscularly with 10^9 (**a, b**), 10^8 (**c, d**), 10^7 (**e, f**) or 10^6 (**g, h**) virus particles of rAd5-Gag (open circles), rAd5HVR48(1)-Gag (filled squares), or rAd5HVR48(1–7)-Gag (filled triangles). **a–h**, Gag epitope-specific CD8⁺ T lymphocyte responses were assessed by D^p/AL11 tetramer binding assays at multiple time points after immunization. **i, j**, Gag-specific cellular immune responses to pooled Gag peptides as well as to AL11, KV9 and DD13 epitope peptides were assessed by IFN- γ ELISPOT assays on day 28 after immunization in mice that received 10^8 virus particles of each vector. Spot-forming cells (SFC) per million splenocytes are shown. **k, l**, Naive mice or mice with anti-Ad5 immunity were immunized intramuscularly with 10^9 or 10^7 virus particles of rAd5-Gag, rAd5HVR48(1–7)-Gag, rAd35-Gag, or rAd48-Gag. Gag epitope-specific CD8⁺ T lymphocyte responses were assessed by D^p/AL11 tetramer binding assays on day 21 after immunization. Error bars are means $\pm$ s.e.m.

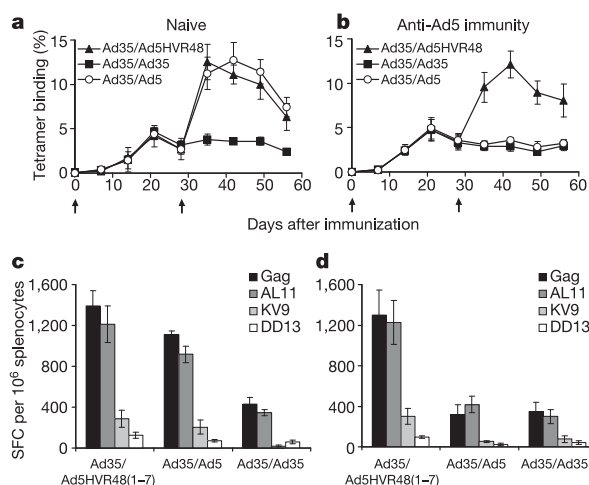


Figure 3 | Immunogenicity of heterologous recombinant adenovirus prime-boost regimens. Naive C57/BL6 mice (**a, c**) or C57/BL6 mice with anti-Ad5 immunity (**b, d**) were primed on day 0 with 10^9 virus particles of rAd35-Gag and boosted on day 28 with 10^9 virus particles of rAd5HVR48(1–7)-Gag, rAd35-Gag or rAd5-Gag. Arrows indicate immunizations. **a, b**, Gag epitope-specific CD8⁺ T lymphocyte responses were assessed by D^p/AL11 tetramer binding assays at multiple time points after immunization. **c, d**, Gag-specific cellular immune responses to pooled Gag peptides as well as to AL11, KV9 and DD13 epitope peptides were assessed by IFN- γ ELISPOT assays on day 56 after the primary immunization. Error bars are means $\pm$ s.e.m.

pilot study comparing the immunogenicity of rAd5-Gag and rAd5HVR48(1–7)-Gag in ten *Mamu-A*01*-negative rhesus monkeys (Fig. 5). Monkeys were pre-immunized twice with 10^{11} virus particles of rAd5-Empty to generate Ad5-specific neutralizing antibody titres of 16,384–32,768. Naive monkeys and monkeys with anti-Ad5 immunity were then immunized intramuscularly with a single injection of 10^{11} virus particles of rAd5-Gag or rAd5HVR48(1–7)-Gag. In naive monkeys ($n = 3$ per group), both vectors elicited potent and comparable Gag-specific IFN- γ ELISPOT responses (Fig. 5a, b). Multiparameter intracellular cytokine staining (ICS) assays¹⁷ also suggested that these responses were comparable and consisted primarily of central memory CD8⁺ T lymphocyte responses (CD3⁺CD8⁺CD28⁺CD95⁺) and effector memory CD8⁺ T lymphocyte responses (CD3⁺CD8⁺CD28⁻CD95⁺), although lower frequencies of CD4⁺ central memory T lymphocyte responses (CD3⁺CD4⁺CD28⁺CD95⁺) were also observed (Fig. 5c, d). In monkeys with anti-Ad5 immunity ($n = 2$ per group), low responses were still detected after rAd5-Gag immunization, presumably as a result of the high dose of the vaccine used in this study. These responses, however, were approximately fourfold lower than those observed in naive monkeys (Fig. 5a, c), consistent with the results of previous studies⁸. In contrast, responses elicited by rAd5HVR48(1–7)-Gag were not detectably suppressed by anti-Ad5 immunity (Fig. 5b, d). Although statistical analyses could not be performed due to the limited number of animals in this pilot study, these data nevertheless show that rAd5HVR48(1–7) vectors are highly immunogenic in nonhuman primates with anti-Ad5 immunity.

These studies demonstrate that targeted HVR modification strategies can be used to construct chimaeric rAd5 vectors that effectively circumvent anti-Ad5 immunity in both mice and rhesus monkeys. Previous reports have shown that exchanging complete hexon genes between viruses from different adenovirus subgroups results in non-viable or poorly viable viruses⁹, presumably due to viral structural constraints. In particular, highly immunogenic adenovirus subgroup C vectors (such as rAd5 vectors) containing hexons from rare serotype adenovirus subgroup D viruses (such as Ad48) have not previously been successfully constructed. We speculate that exchanging only the seven short HVRs on the surface of the Ad5

hexon protein effectively preserves the hexon core structure while removing key neutralizing determinants. The HVR-chimaeric rAd5 vectors also preserve many favourable features of rAd5 vectors, including vector stability, specific infectivity, high levels of transgene expression and potent immunogenicity. Production yields of the rAd5HVR48(1–7) vectors, however, remained several-fold lower than rAd5 vectors, suggesting that these vectors can be further optimized. It should also be possible to construct a series of HVR-modified rAd5 vectors with HVRs derived from other adenovirus serotypes as well as with synthetic sequences to facilitate heterologous recombinant adenovirus prime-boost vaccine regimens, although the extent of HVR sequence constraints in these chimaeric vectors remains to be determined.

The observations that rAd5HVR48(1–7) vectors exhibit primarily the serologic properties of Ad48 rather than Ad5 *in vitro* and effectively circumvent anti-Ad5 immunity *in vivo* demonstrate that dominant Ad5-specific neutralizing antibodies are focused primarily on epitopes within the hexon HVRs. We estimate that 10–15% of Ad5-specific neutralizing antibodies are directed against other Ad5 capsid proteins, including fibre and penton^{18,19}, based on virus neutralization studies using murine and human serum samples (Fig. 4). Fibre- and penton-specific neutralizing antibodies, however, are probably less relevant than hexon-specific neutralizing antibodies, as they are typically lower in titre and also seem inefficient at suppressing rAd5 vaccine immunogenicity *in vivo*⁶, perhaps reflecting different mechanisms of virus neutralization²⁰. Our conclusion that functionally relevant Ad5-specific neutralizing antibodies are focused primarily on the hexon HVRs is also consistent with studies that have shown that these solvent-exposed loops comprise the majority of adenovirus serotype-specific sequence variability^{11,13}. We cannot, however, exclude the possibility that Ad5-specific neutralizing antibodies against other epitopes as well as Ad5-specific cellular immune responses may have important secondary roles in certain settings.

These studies have major implications for the development of recombinant viral vectors for vaccination and gene therapy. The identification and specific removal of dominant neutralizing antibody epitopes on the surface of viral capsid proteins offers a novel strategy to create viral vectors that evade pre-existing neutralizing antibody responses and that overcome the critical limitation of anti-vector immunity. In particular, HVR-chimaeric rAd5 vectors may prove useful as vaccine vectors for HIV-1 and other pathogens that

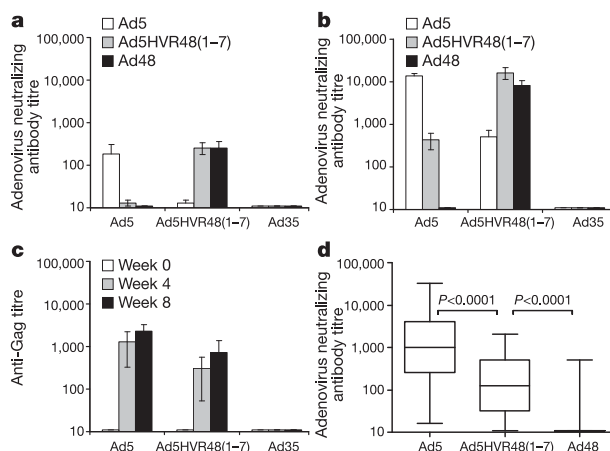


Figure 4 | Humoral immune responses to hexon HVR-chimaeric rAd5 vectors. **a–c**, Naive mice were primed at week 0 and boosted at week 4 with 10^{10} virus particles of rAd5-Gag, rAd5HVR48(1–7)-Gag, or rAd55-Gag. **a, b**, Ad5-specific (white bars), Ad5HVR48(1–7)-specific (grey bars) and Ad48-specific (black bars) neutralizing antibody titres were assessed by virus neutralization assays using serum samples obtained at week 4 (**a**) and week 8 (**b**). **c**, Gag-specific antibodies were assessed at weeks 0 (white bars), 4 (grey bars) and 8 (black bars) by ELISA. **d**, Ad5-specific, Ad5HVR48(1–7)-specific and Ad48-specific neutralizing antibody titres were assessed in 265 serum samples from South Africa and are presented as box-and-whisker plots (see Methods). Error bars are means $\pm$ s.e.m.

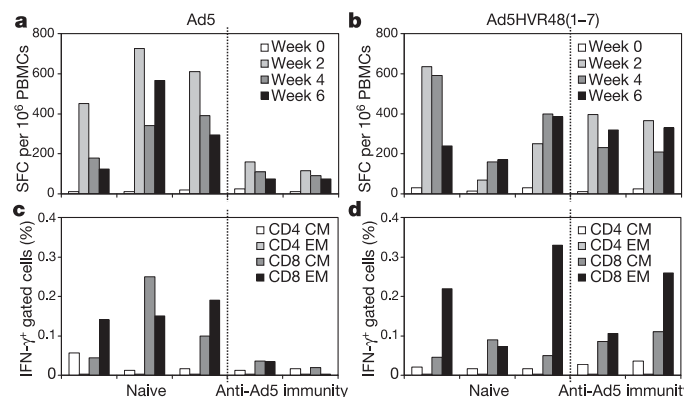


Figure 5 | Immunogenicity of hexon HVR-chimaeric rAd5 vectors in rhesus monkeys. Naive rhesus monkeys or rhesus monkeys with anti-Ad5 immunity were immunized intramuscularly with 10^{11} virus particles of rAd5-Gag (**a, c**) or rAd5HVR48(1–7)-Gag (**b, d**). **a, b**, Gag-specific cellular immune responses were assessed by IFN- γ ELISPOT assays at multiple time points after immunization. **c, d**, Gag-specific CD4⁺ and CD8⁺ central memory (CM; CD28⁺CD95⁺) and effector memory (EM; CD28⁻CD95⁺) T lymphocyte responses were evaluated by pooled peptide IFN- γ ICS assays at week 4 after immunization.

are endemic in the developing world, where the vast majority of individuals have high levels of pre-existing anti-Ad5 immunity.

METHODS

Vector cloning and production. E1/E3-deleted, replication-incompetent rAd5 vectors containing chimaeric hexon genes were constructed using plasmid/cosmid recombination systems as described^{4,21}. Partial Ad5 hexon genes containing the Ad5 HVRs exchanged with the corresponding regions from Ad48 were produced synthetically (GeneART) and cloned as *ApaI*-*HpaI* fragments into a shuttle plasmid containing the complete Ad5 hexon gene. The Ad5 HVR regions were defined as amino acids 136–165 (HVR1), 188–194 (HVR2), 212–220 (HVR3), 248–258 (HVR4), 268–281 (HVR5), 305–310 (HVR6) and 418–451 (HVR7). *AscI*-*AscI* fragments containing the complete chimaeric hexon genes were then excised from the shuttle plasmids and used to replace the corresponding regions in the Ad5 cosmid pWEAd5.AflIII-rITR.dE3. The resultant mutant Ad5 cosmids together with the adaptor plasmid pAdApt expressing codon-optimized SIVmac239 Gag under control of a CMV promoter were co-transfected into complementing PER.C6/55K cells, and homologous recombination yielded rAd5HVR48(1)–Gag and rAd5HVR48(1–7)–Gag vectors. These vectors were plaque-purified, expanded and purified by CsCl gradient centrifugation. Replication-incompetent rAd48 vectors were constructed using methods we have described for other serotypes⁴ and exhibited comparable expression and yields.

Animals and immunizations. Six-to-eight-week-old C57/BL6 mice (Charles River Laboratories) were injected intramuscularly with varying doses of replication-incompetent recombinant adenovirus vectors expressing SIV Gag in 100 µl sterile PBS divided equally in both quadriceps muscles. To induce active anti-vector immunity, mice were pre-immunized intramuscularly twice, separated by a 4-week interval, with 10¹⁰ virus particles of rAd5-Empty in 100 µl sterile PBS. Adult outbred rhesus monkeys that did not express the *Mamu-A*01* class I allele were injected intramuscularly with 10¹¹ virus particles of recombinant adenovirus vectors expressing SIV Gag in 1 ml sterile PBS divided equally in both quadriceps muscles. To induce active anti-vector immunity, monkeys were pre-immunized intramuscularly twice, separated by an 8-week interval, with 10¹¹ virus particles rAd5-Empty in 1 ml sterile PBS.

Tetramer binding assays. Tetrameric H-2D^b complexes folded around the immunodominant SIV Gag AL11 epitope (AAVKNWMTQTL) were prepared and used to stain epitope-specific CD8⁺ T lymphocytes from C57/BL6 mice as described⁷. Samples were analysed by two-colour flow cytometry on a FACS Array (BD Pharmingen), and gated CD8⁺ T lymphocytes were examined for staining with the D^b/AL11 tetramer. CD8⁺ T lymphocytes from naive mice exhibited <0.1% tetramer staining.

ELISPOT assays. SIV Gag-specific cellular immune responses in splenocytes from vaccinated mice and peripheral blood mononuclear cells (PBMCs) from vaccinated monkeys were assessed by IFN-γ ELISPOT assays in response to overlapping 15-amino-acid peptides spanning the SIVmac239 Gag protein (NIH AIDS Research and Reference Reagent Program) as well as epitope peptides as described^{6,7,22}.

ICS assays. SIV Gag-specific CD4⁺ and CD8⁺ T lymphocyte responses in PBMCs from vaccinated monkeys were assessed by multiparameter ICS assays as described¹⁷. Cells were stained with monoclonal antibodies against CD4-FITC (L200), CD95-PE (DX2), CD28-PerCP-Cy5.5 (L293), IFN-γ-PE-Cy7 (B27), IL-2-APC (MQ1-17H12), CD3-Alexa700 (SP34) and CD8-APC-Cy7 (SK1) (BD Biosciences).

Virus neutralization assays. Ad5-specific, Ad5HVR48(1–7)-specific and Ad48-specific neutralizing antibody titres in murine and human serum samples were measured by luciferase-based virus neutralization assays as described¹⁶. Serum samples from healthy individuals randomly selected from a South African Medical Research Council measles vaccine study were used for the human adenovirus seroprevalence studies. Neutralization titres were defined as the maximum serum dilution that neutralized 90% of luciferase activity.

ELISAs. Gag-specific antibody responses in vaccinated mice were assessed by a direct ELISA using purified recombinant SIV Gag protein (Intracel) as described⁷.

Statistical analyses. Statistical analyses were performed with GraphPad Prism version 4.01 (GraphPad Software). Immune response data are presented as means with standard errors. Comparisons of mean immune responses were performed by two-tailed ANOVA with Bonferroni adjustments to account for multiple comparisons. Neutralizing antibody data using human serum samples are presented as box-and-whisker plots depicting medians, 25–75% interquartile

ranges and complete ranges. Comparisons of median neutralizing antibody titres were performed by two-tailed Wilcoxon rank-sum tests.

Received 12 January; accepted 15 March 2006.

Published online 16 April 2006.

- Shiver, J. W. *et al.* Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. *Nature* **415**, 331–335 (2002).
- Sullivan, N. J. *et al.* Accelerated vaccination for Ebola virus haemorrhagic fever in non-human primates. *Nature* **424**, 681–684 (2003).
- Shiver, J. W. & Emini, E. A. Recent advances in the development of HIV-1 vaccines using replication-incompetent adenovirus vectors. *Annu. Rev. Med.* **55**, 355–372 (2004).
- Vogels, R. *et al.* Replication-deficient human adenovirus type 35 vectors for gene transfer and vaccination: efficient human cell interaction and bypass of preexisting adenovirus immunity. *J. Virol.* **77**, 8263–8271 (2003).
- Kostense, S. *et al.* Adenovirus types 5 and 35 seroprevalence in AIDS risk groups supports type 35 as a vaccine vector. *AIDS* **18**, 1213–1216 (2004).
- Sumida, S. M. *et al.* Neutralizing antibodies to adenovirus serotype 5 vaccine vectors are directed primarily against the adenovirus hexon protein. *J. Immunol.* **174**, 7179–7185 (2005).
- Barouch, D. H. *et al.* Immunogenicity of recombinant adenovirus serotype 35 vaccine in the presence of pre-existing anti-Ad5 immunity. *J. Immunol.* **172**, 6290–6297 (2004).
- Casimiro, D. R. *et al.* Comparative immunogenicity in rhesus monkeys of DNA plasmid, recombinant vaccinia virus, and replication-defective adenovirus vectors expressing a human immunodeficiency virus type 1 gag gene. *J. Virol.* **77**, 6305–6313 (2003).
- Youil, R. *et al.* Hexon gene switch strategy for the generation of chimeric recombinant adenovirus. *Hum. Gene Ther.* **13**, 311–320 (2002).
- Gall, J. G., Crystal, R. G. & Falck-Pedersen, E. Construction and characterization of hexon-chimeric adenoviruses: specification of adenovirus serotype. *J. Virol.* **72**, 10260–10264 (1998).
- Crawford-Miksza, L. & Schnurr, D. P. Analysis of 15 adenovirus hexon proteins reveals the location and structure of seven hypervariable regions containing serotype-specific residues. *J. Virol.* **70**, 1836–1844 (1996).
- Athapilly, F. K., Murali, R., Rux, J. J., Cai, Z. & Burnett, R. M. The refined crystal structure of hexon, the major coat protein of adenovirus type 2, at 2.9 Å resolution. *J. Mol. Biol.* **242**, 430–455 (1994).
- Rux, J. J., Kuser, P. R. & Burnett, R. M. Structural and phylogenetic analysis of adenovirus hexons by use of high-resolution x-ray crystallographic, molecular modeling, and sequence-based methods. *J. Virol.* **77**, 9553–9566 (2003).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Kraulis, P. J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Sprangers, M. C. *et al.* Quantifying adenovirus-neutralizing antibodies by luciferase transgene detection: addressing preexisting immunity to vaccine and gene therapy vectors. *J. Clin. Microbiol.* **41**, 5046–5052 (2003).
- Pitcher, C. J. *et al.* Development and homeostasis of T cell memory in rhesus macaque. *J. Immunol.* **168**, 29–43 (2002).
- Gahery-Segard, H. *et al.* Immune response to recombinant capsid proteins of adenovirus in humans: antifiber and anti-penton base antibodies have a synergistic effect on neutralizing activity. *J. Virol.* **72**, 2388–2397 (1998).
- Hong, S. S., Habib, N. A., Franqueville, L., Jensen, S. & Boulanger, P. A. Identification of adenovirus (Ad) penton base neutralizing epitopes by use of sera from patients who had received conditionally replicative Ad (Ad1520) for treatment of liver tumors. *J. Virol.* **77**, 10366–10375 (2003).
- Wohlfart, C. Neutralization of adenoviruses: kinetics, stoichiometry, and mechanisms. *J. Virol.* **62**, 2321–2328 (1988).
- Havenga, M. J. *et al.* Improved adenovirus vectors for infection of cardiovascular tissues. *J. Virol.* **75**, 3335–3342 (2001).
- Nanda, A. *et al.* Immunogenicity of recombinant fiber-chimeric adenovirus serotype 35 vector-based vaccines in mice and rhesus monkeys. *J. Virol.* **79**, 14161–14168 (2005).

Acknowledgements We thank R. Vogels, J. Custers, N. Letvin, R. Dolin, A. Kelcz, Y. Sun, L. Shen, M. Kishko, D. Truitt, S. Harrison, B. Walker and P. Kiepiela for advice, assistance and reagents. The peptide pools were obtained from the NIH AIDS Research and Reference Reagent Program. We acknowledge support from NIH grants to D.H.B.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.H.B. (dbarouch@bidmc.harvard.edu).

LETTERS

A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting

Olivier Namy^{1*†}, Stephen J. Moran^{2*}, David I. Stuart^{2,3}, Robert J. C. Gilbert^{2,3} & Ian Brierley¹

The triplet-based genetic code requires that translating ribosomes maintain the reading frame of a messenger RNA faithfully to ensure correct protein synthesis¹. However, in programmed -1 ribosomal frameshifting², a specific subversion of frame maintenance takes place, wherein the ribosome is forced to shift one nucleotide backwards into an overlapping reading frame and to translate an entirely new sequence of amino acids. This process is indispensable in the replication of numerous viral pathogens, including HIV and the coronavirus associated with severe acute respiratory syndrome³, and is also exploited in the expression of several cellular genes⁴. Frameshifting is promoted by an mRNA signal composed of two essential elements: a heptanucleotide 'slippery' sequence⁵ and an adjacent mRNA secondary structure, most often an mRNA pseudoknot⁶. How these components operate together to manipulate the ribosome is unknown. Here we describe the observation of a ribosome–mRNA pseudoknot complex that is stalled in the process of -1 frameshifting. Cryoelectron microscopic imaging of purified mammalian 80S ribosomes from rabbit reticulocytes paused at a coronavirus pseudoknot reveals an intermediate of the frameshifting process. From this it can be seen how the pseudoknot interacts with the ribosome to block the mRNA entrance channel, compromising the translocation process and leading to a spring-like deformation of the P-site transfer RNA. In addition, we identify movements of the likely eukaryotic ribosomal helicase and confirm a direct interaction between the translocase eEF2 and the P-site tRNA. Together, the structural changes provide a mechanical explanation of how the pseudoknot manipulates the ribosome into a different reading frame.

Programmed -1 ribosomal frameshifting was first described in 1985 (ref. 7) but the mechanism of the process has remained elusive. Subsequently^{8–10} it was shown that RNA pseudoknot structures pause ribosomes over a homopolymeric slippery sequence, at which the ribosome-bound tRNAs realign in the -1 frame⁵, but that pausing alone is insufficient to promote frameshifting^{8–11}. To gain insight into this mechanism at the molecular level, we developed a method of purifying rabbit reticulocyte lysate (RRL) ribosomes stalled in the act of decoding a frameshift signal and subjected them to cryoelectron microscopy (cryo-EM) analysis (Methods and Supplementary Fig. 1). Stalled complexes were assembled on a short mRNA transcript containing a variant of the coronavirus IBV frameshift signal⁶ in which the slippery sequence was replaced by a sequence that is unslippery¹² (CGAGGCA). Such complexes have been shown to represent true intermediates, and not 'dead-end' products^{9–11}. Three reconstructions of RRL ribosomes were produced. The first was the translating 80S ribosome, stalled at the IBV pseudoknot (PK) and containing densities consistent with the presence of a P-site

tRNA and elongation factor 2 (eEF2) as judged from modelling and comparative analysis (80S_{PK}; Fig. 1a). The resolution of this reconstruction was 16.2 Å (see Supplementary Information), for which 12,161 out of a total of 17,672 images were selected as being stalled with full and equal occupancy of eEF2, tRNA and pseudoknot. Structural changes associated with pseudoknot-induced -1 frameshifting were then determined by reference to the two control reconstructions. The first of these was an apo-80S ribosome (80S_{Apo}; Fig. 1b), incorporating 10,296 images with a resolution of 14 Å. The second was created by deletion of the pseudoknot loops and connection of the two stems into a single stem-loop (SL; Supplementary Fig. 1) to form a structure capable of stalling the ribosome sufficiently to allow purification but leaving it unable to promote efficient frameshifting¹¹ (80S_{SL}; Fig. 1c). The 80S_{SL} reconstruction resolution was 15.7 Å, for which 9,805 out of 14,887 images were selected as being stalled with tRNA bound. The ribosomes purified with this stem-loop were treated and prepared in exactly the same way as the pseudoknot-engaged ribosome. Any differences observed between these two reconstructions, such as bound cofactors, can therefore be ascribed directly to the action of the functional pseudoknot.

Our reconstructions reveal that the pseudoknot interacts with the ribosome at the entrance to the mRNA channel, in close association with a likely mammalian 80S helicase¹³ (Fig. 2, left panel). This confirms previous suggestions^{13,14} of the possible position of the pseudoknot structure. From the path taken by a short synthetic mRNA through the 70S ribosome decoding site, it has been estimated that mRNA enters the ribosome around nucleotides +13 to +15 (where the first residue of the P-site AUG codon is designated +1 (ref. 14)). In our mRNA the first base of the pseudoknot was at position +14 with respect to the first base of the P-site and thus would be expected to be located at the entrance to the mRNA tunnel as we see it in our map. It has also been shown¹³ that the 70S ribosome can itself act as a helicase for unwinding mRNA secondary structures, with the active site at position +11, located between the head and shoulder of the 30S subunit. Prokaryotic ribosomal proteins S3, S4 and S5 that line the entrance to the tunnel are implicated in helicase activity; they potentially form a ring around the incoming mRNA, acting as a processivity clamp. Superimposition of our pseudoknot-stalled and control maps allows a visual representation of differences between the maps found in this region (Fig. 2). It has been suggested that the mode of action of the pseudoknot arises from stereochemical mismatch between the ribosomal helicase and the pseudoknot structure¹⁴. In our pseudoknot-engaged map, three regions of density surrounding the mRNA entrance tunnel seem to move up and contact the putative pseudoknot structure (Fig. 2, left panel). A yeast model¹⁵ identifies one of these regions as rpS3 (in prokaryotes,

¹Division of Virology, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK. ²Division of Structural Biology, Henry Wellcome Building for Genomic Medicine, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. ³Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QH, UK. [†]Present address: Institut de Génétique et Microbiologie, UMR8621, Université Paris-Sud, 91405 Orsay, France.

*These authors contributed equally to this work.

S3), one as 18S rRNA helix 16, which interacts with rpS9 (S4 in prokaryotes), and a final region as rpS2 (S5 in prokaryotes). In addition, the ubiquitous eukaryotic ribosomal regulatory protein RACK1 undergoes a movement that seems connected to the presence of the pseudoknot. This observation represents the first view of a ribosomal helicase in action. Furthermore, the interaction of the pseudoknot with the mammalian equivalents of S3, S4 and S5 indicates that part of the pseudoknot's functional significance for the elongating ribosome is indeed an interaction with the ribosomal helicase. Resistance of the pseudoknot to the action of the helicase may contribute to the stability of the overall complex, trapping the elongating ribosome.

The 80S_{PK} complex shows a ratchet-like rearrangement of the ribosome linked to the trapping of eEF2 (refs 16, 17). No antibiotics were used to trap eEF2 (as were used in previous studies), indicating that such a structure might indeed be physiological¹⁸. As neither eEF2 nor a ratchet-like rotation is observed in the 80S_{SL} complex, in the 80S_{PK}, the changes can be attributed to the functional pseudoknot structure. These features indicate that the 80S_{PK} ribosome has been stalled by the pseudoknot during translocation, with eEF2 precluding the binding of a tRNA in the A-site until the frameshift has been completed. This first visualization of an elongating eukaryotic ribosome stalled in the act of translocation confirms a direct interaction between eEF2, the translocase, and the P-site tRNA. We can therefore identify translocation as the point in the elongation cycle at which frameshifting occurs, as has been suggested previously^{19,20}. This interplay of pseudoknot, tRNA, eEF2 and structural rearrangement of the ribosome directly illustrates the dynamic and concerted nature of the frameshifting process.

The P-site tRNA in the pseudoknot-engaged reconstruction is distorted in comparison with the stem-loop reconstruction (Fig. 2, right panel) such that the anticodon is raised by a bending of the tRNA towards the A-site occupied by the tip of eEF2. Although the shape of the T-arm and acceptor stem remains unmodified, the elbow of the tRNA is pushed upwards towards the roof of the P-site (against the large subunit), indicating the presence of opposing forces. The present resolution (16 Å) is insufficient to permit the visualization of atomic details; however, it seems that the conformational change in the tRNA derives essentially from a bending of the D-arm (Fig. 2,

right panel). The distortion of the P-site tRNA in these complexes is similar to the flexibility of the amino-acyl tRNA previously observed during the accommodation process²¹, in which the processes of codon–anticodon interaction and accommodation of cognate tRNAs are viewed²¹ as a dynamic interplay involving the mechanical properties of tRNA as well as those of the ribosome. The similar mechanical deformation observed in our pseudoknot-stalled complex indicates that the spring-like properties of tRNA might also be important in frameshifting.

The insight into the frameshifting process presented here leads us to propose a mechanical explanation for programmed –1 frameshifting (Fig. 3). When the elongating ribosome encounters the pseudoknot structure, the helicase action at the mRNA entrance tunnel attempts to unwind the pseudoknot but stereochemical mismatch prevents it from doing so. During translocation, the movement of tRNA through the ribosome is resisted by tension developed in the mRNA strand by the pseudoknot blockage at the tunnel entrance (as suggested previously²²). The mRNA in turn is connected to the tRNA by means of the codon–anticodon interaction. Because the tRNA is prevented from returning to the A-site by the presence of eEF2, the ribosome, in attempting to translocate the anticodon into the authentic P-site, places strain on the tRNA that results in the adoption of a bent conformation. The opposing actions of translocation, catalysed by eEF2, and pulling from the mRNA strand account for the bending of the tRNA, spring-like, in a (+) sense (3' direction) and the movement of the elbow of the tRNA into the roof of the P-site. These opposing forces place a strain on the codon–anticodon interaction that promotes breakage. Subsequent relaxation of the bent tRNA structure would then be in a (–) sense

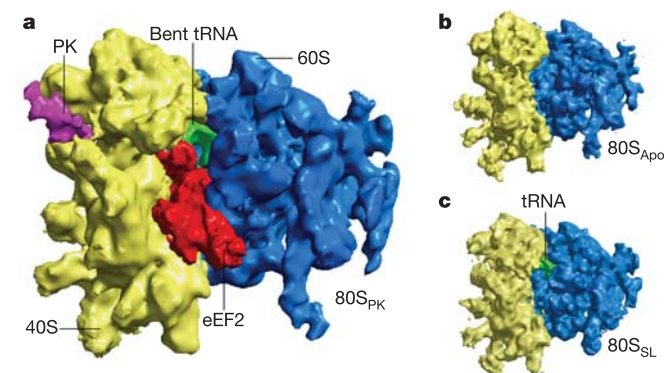


Figure 1 | Structures of stalled 80S complexes. **a**, The pseudoknot-engaged rabbit 80S ribosome (80S_{PK}) viewed in profile. The large 60S subunit is coloured blue and the small 40S subunit is coloured yellow. The P-site tRNA stalled within the complex is coloured green, the eukaryotic translocase (eEF2) red and the pseudoknot structure (PK) purple. **b**, The control apo-80S ribosome (80S_{Apo}) with the subunits coloured as in **b**. **c**, The structure of a ribosome purified in the same way as 80S_{PK}, but with the pseudoknot modified to form a stem-loop with greatly reduced capacity to induce frameshifting (80S_{SL}). The subunits are coloured as in **b** and **c**; the P-site tRNA stalled within the complex is coloured green.

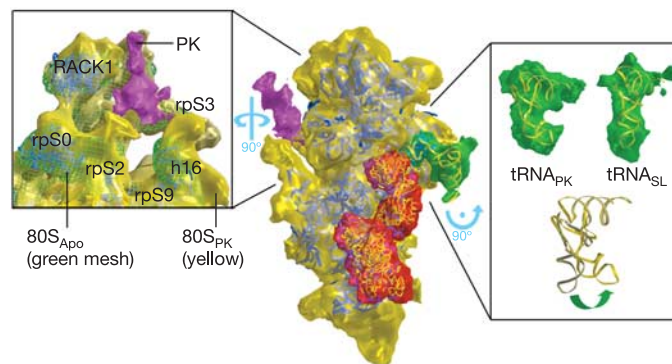


Figure 2 | Atomic fits to the small subunit and bound cofactors of the pseudoknot-engaged ribosome. In the centre the small subunit is viewed in a similar orientation to that of Fig. 1, with the large subunit removed computationally. The subunit, tRNA, eEF2 and pseudoknot are coloured as in Fig. 1. The yeast atomic model for the small subunit¹⁵ has been fitted to the subunit itself (blue ribbons), the structure of eEF2 (ref. 18) to the corresponding density (yellow coil), and the tRNA stalled within the complex likewise (yellow ribbon). A stereo view of the central image is provided in Supplementary Fig. 2. Left and right: close-up views, rotated as indicated to afford a detailed picture of the structural changes associated with the engaged pseudoknot (left) and the stalled tRNA (right). Left: a view from the solvent face of the 80S_{PK} small subunit (yellow), with the 80S_{Apo} structure (green mesh) superimposed. Movements of the subunit in the 80S_{PK} ribosome up and towards the pseudoknot structure relative to the 80S_{Apo} ribosome can be seen. Here the atomic fits (blue ribbon) are to the 80S_{Apo} structure. From this fitting the movements can be associated with the eukaryotic equivalents of the prokaryotic helicase. Although the positioning of the pseudoknot itself is clear, the mRNA passing through the entrance channel is too thin to be distinguished at the current resolution. Right: a comparison of the 80S_{PK} (top left) and the 80S_{SL} (top right) tRNAs (density in green; fitted atomic models in yellow). At the bottom a superposition of the two models is shown, showing the bending of the 80S_{PK} tRNA.

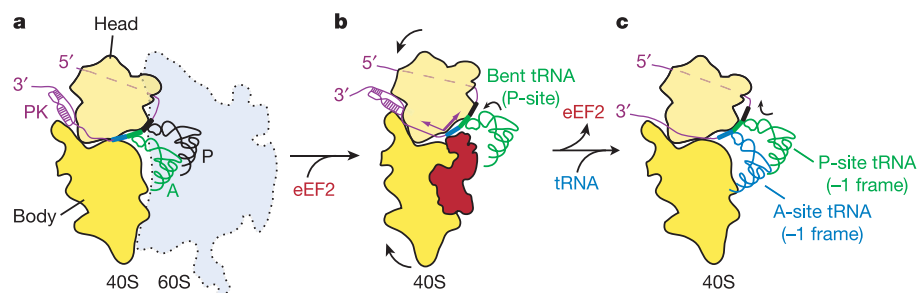


Figure 3 | A mechanical model for pseudoknot-induced -1 frameshifting. Three different states of the small subunit translating an mRNA containing a pseudoknot that induces -1 frameshifting are shown. **a**, The elongating ribosome approaching the pseudoknot in the zero reading frame. **b**, Engagement with the pseudoknot, generating a frameshifting intermediate in which the small subunit is stalled during translocation with eEF2 bound,

causing tension in the mRNA that bends the P-site tRNA in a (+) sense direction. As a result the anticodon-codon interaction breaks over the slippery sequence, allowing a spring-like relaxation of the tRNA in a (-) sense direction. **c**, Re-engagement of the tRNA with the mRNA, leaving the ribosome translating in the -1 reading frame.

(5' direction), allowing the tRNA to re-pair with the mRNA in the -1 position. In support of this model, slippery sequences possessing a G-triplet or C-triplet in the A-site position show greatly reduced frameshifting⁵, which is explicable in terms of increased stability of the codon-anticodon complex and the consequent increased resistance to the mechanical forces applied. The mechanical model also offers an explanation for the capacity of RNAs of differing secondary structure to promote frameshifting to different extents^{2,12}. The interaction between these RNA structures and the 80S helicase is likely to be influenced both by conformational features and relative stability, so the amount of the strain on the tRNA and the likelihood of breakage of the codon-anticodon pair will depend on the secondary structure of the mRNA. For secondary structures that do not lead to frameshifting, including the SL used here, we propose that the helicase unwinds them in defined triplet steps coupled to normal translocation¹³, limiting mRNA tension. Such structures could still induce ribosomal pausing^{10,11}, by virtue of their stability, but would be unable to generate the mRNA tension required for the distortion of tRNA and the induction of frameshifting during translocation.

METHODS

Detailed methods are given in Supplementary Methods.

Isolation of paused 80S mRNA complexes. Messenger RNAs (about 190 nucleotides) containing the minimal IBV pseudoknot²³ or related stem-loop structure²³ were preannealed to a biotinylated RNA oligonucleotide and translated in rabbit reticulocyte lysates for 15 min at 27°C before the addition of cycloheximide. Ribosomes were pelleted through a sucrose cushion, resuspended and loaded on an avidin column. After extensive washing, mRNA-ribosome complexes were released by targeting a DNA oligonucleotide to a region downstream of the pseudoknot/stem-loop and adding RNase H (see Supplementary Fig. 1). Aliquots of ribosomes were taken and frozen at -70°C before cryo-EM.

Cryo-EM, image processing and fitting of atomic structures. After particle picking and correction for contrast transfer function, the 80S_{Apo} data were used to generate a reconstruction *ab initio* with the use of IMAGIC²⁴, refined iteratively in SPIDER²⁵. The 80S_{Apo} map was used to align the 80S_{PK} or 80S_{SL} images, which were then also refined iteratively. Final maps were scaled in reciprocal space and sharpened by the application of a *B* factor of -500 Å². Fourier shell correlation (0.5 criterion) indicated the resolutions of the maps as follows: 80S_{Apo}, 14.0 Å; 80S_{SL}, 15.7 Å; 80S_{PK}, 16.2 Å. Atomic fitting was performed with O²⁶ and URO²⁷, making use of the published yeast atomic model¹⁵, and Protein Data Bank entries 1NOU and 2TRA. Figures were generated with BOBSCRIPT²⁸ and Raster3D²⁹.

Received 3 February; accepted 20 March 2006.

- Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
- Farabaugh, P. J. Programmed translational frameshifting. *Microbiol. Rev.* **60**, 103–134 (1996).

- Brierley, I. & Dos Ramos, F. Ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res.* advance online publication, 28 November 2005 (doi:10.1016/j.virusres.2005.10.008).
- Namy, O., Rousset, J.-P., Naphine, S. & Brierley, I. Reprogrammed genetic decoding in cellular gene expression. *Mol. Cell* **13**, 157–168 (2004).
- Jacks, T., Madhani, H. D., Masiarz, F. R. & Varmus, H. E. Signals for ribosomal frameshifting in the Rous sarcoma virus *gag-pol* region. *Cell* **55**, 447–458 (1988).
- Brierley, I., Digard, P. & Inglis, S. C. Characterisation of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell* **57**, 537–547 (1989).
- Jacks, T. & Varmus, H. E. Expression of the Rous sarcoma virus *pol* gene by ribosomal frameshifting. *Science* **230**, 1237–1242 (1985).
- Tu, C., Tzeng, T.-H. & Bruenn, J. A. Ribosomal movement impeded at a pseudoknot required for frameshifting. *Proc. Natl Acad. Sci. USA* **89**, 8636–8640 (1992).
- Lopinski, J. D., Dinman, J. D. & Bruenn, J. A. Kinetics of ribosomal pausing during programmed -1 ribosomal frameshifting. *Mol. Cell. Biol.* **20**, 1095–1103 (2000).
- Kontos, H., Naphine, S. & Brierley, I. Ribosomal pausing at a frameshift RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. *Mol. Cell. Biol.* **21**, 8657–8670 (2001).
- Somogyi, P., Jenner, A. J., Brierley, I. & Inglis, S. C. Ribosomal pausing during translation of an RNA pseudoknot. *Mol. Cell. Biol.* **13**, 6931–6940 (1993).
- Atkins, J. F. *et al.* Overriding standard decoding: Implications of recoding for ribosome function and enrichment of gene expression. *Cold Spring Harb. Symp. Quant. Biol.* **66**, 217–232 (2001).
- Takyar, S., Hickerson, R. P. & Noller, H. F. mRNA helicase activity of the ribosome. *Cell* **120**, 49–58 (2005).
- Yusupova, G. Z., Yusupov, M. M., Cate, J. H. & Noller, H. F. The path of messenger RNA through the ribosome. *Cell* **106**, 233–241 (2001).
- Spahn, C. M. *et al.* Structure of the 80S ribosome from *Saccharomyces cerevisiae*: tRNA-ribosome and subunit-subunit interactions. *Cell* **107**, 373–386 (2001).
- Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406**, 318–322 (2000).
- Spahn, C. M. *et al.* Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. *EMBO J.* **23**, 1008–1019 (2004).
- Jorgensen, R. *et al.* Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase. *Nature Struct. Biol.* **10**, 379–385 (2003).
- Weiss, R. B., Dunn, D. M., Shuh, M., Atkins, J. F. & Gesteland, R. F. *E. coli* ribosomes re-phase on retroviral frameshift signals at rates ranging from 2 to 50 percent. *New Biol.* **1**, 159–169 (1989).
- Yelverton, E., Lindsley, D., Yamauchi, P. & Gallant, J. A. The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in *Escherichia coli*. *Mol. Microbiol.* **11**, 303–313 (1994).
- Valle, M. *et al.* Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. *Nature Struct. Biol.* **10**, 899–906 (2003).
- Plant, E. P. *et al.* The 9 Å solution: How mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. *RNA* **9**, 168–174 (2003).
- Brierley, I., Jenner, A. J. & Inglis, S. C. Mutational analysis of the 'slippery sequence' component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.* **227**, 463–479 (1992).
- van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).
- Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D

- electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
26. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
 27. Navaza, J., Lepault, J., Rey, F. A., Alvarez-Rua, C. & Borge, J. On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. *Acta Crystallogr. D* **58**, 1820–1825 (2002).
 28. Esnouf, R. M. Further additions to MolScript version 1.4, including reading and contouring of electron-density maps. *Acta Crystallogr. D* **55**, 938–940 (1999).
 29. Merritt, E. A. & Murphy, M. E. Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I.B. thanks P. Farabaugh for helpful discussions. This work was supported by grants from the Biotechnology and Biological Sciences

Research Council, UK, and the Medical Research Council, UK. S.J.M. holds a Wellcome Trust Structural Biology Studentship, D.I.S. is an MRC Research Professor, and R.J.C.G. is a Royal Society University Research Fellow.

Author Contributions O.N. and S.J.M. contributed equally to this work. O.N. and I.B. purified the stalled ribosome complexes; S.J.M., D.I.S. and R.J.C.G. solved the structures. All authors discussed the results and contributed to writing of the manuscript.

Author Information Electron-density maps have been deposited in the European Bioinformatics Institute Electron Microscopy database, accession numbers EMD-1197, EMD-1198 and EMD-1199 (www.ebi.ac.uk/msd/iims/3D_EMdep.html). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.B. (ib103@mole.bio.cam.ac.uk) or R.J.C.G. (gilbert@strubi.ox.ac.uk).

CORRIGENDUM

doi:10.1038/nature04775

Lipid-protein interactions in double-layered two-dimensional AQP0 crystals

Tamir Gonen, Yifan Cheng, Piotr Sliz, Yoko Hiroaki, Yoshinori Fujiyoshi, Stephen C. Harrison & Thomas Walz

Nature 438, 633–638 (2005)

We wish to clarify that we calculated the *G* factor of -1.5 with the initially deposited coordinates for non-junctional AQP0, 1TM8, and not with 1YMG as stated in this Article, which replaced 1TM8 (January 2005) and which has a *G* factor of 0.3. Also, we did not deposit the structure factors for the X-ray structure of non-junctional AQP0 at the Protein Data Bank, as indicated in the Author Information section: in fact, we deposited our alternative modelling of non-junctional AQP0 (PDB accession number is 2B6P).

CORRIGENDUM

doi:10.1038/nature04777

An acidic protein aligns magnetosomes along a filamentous structure in magnetotactic bacteria

André Scheffél, Manuela Gruska, Damien Faivre, Alexandros Linaroudis, Peter L. Graumann, Jürgen M. Plitzko & Dirk Schüler

Nature 440, 110–114 (2006)

Peter L. Graumann was accidentally omitted from the published author list. His affiliation is Universität Freiburg, Institute for Microbiology, Faculty for Biology, Schänzlestrasse 1, 79104 Freiburg, Germany.

CORRIGENDUM

doi:10.1038/nature04776

Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee

Jennifer F. Hughes, Helen Skaletsky, Tatyana Pyntikova, Patrick J. Minx, Tina Graves, Steve Rozen, Richard K. Wilson & David C. Page

Nature 437, 101–104 (2005)

There is an error in the analysis of one of the genes described in this Letter. The open reading frame (ORF)-truncating frameshift in chimpanzee *CYorf15A* reported in Table 1 is incorrect: we now find that both human and chimpanzee *CYorf15A* genes have ORFs of the same length, so chimpanzee *CYorf15A* does not have a truncated ORF relative to its human orthologue. This finding does not affect our conclusions with respect to preservation of human Y genes. It does mean, however, that the chimpanzee Y chromosome has lost four genes, rather than five genes as stated.

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Good in parts

A report card on attempts to help US postdocs gives mixed marks for research institutions trying to improve the plight of their fellows.

In 2003, only 12 of the 60 institutions surveyed by the National Postdoctoral Association (NPA) had both postdoc offices and postdoc associations. By 2006, when 120 institutions were surveyed, the number of postdoc offices had ballooned to 87, postdoc associations numbered 68, and 47 institutions had both.

This growth is important, because postdoc offices provide a formal mechanism for institutions to address postdoc concerns, and associations give fellows a unified voice to push for reforms on stipends, benefits and career information.

However, grades were less good in another category on the report card: creating long-term tracking of postdoc outcomes. Following where postdocs go when they leave their institutions is helpful, because it gives current fellows an idea of the jobs they can expect. And postdoc alumni databases could also provide a network of speakers for those wishing to hear about non-traditional paths, says

Alyson Reed, executive director of the NPA, based in Washington DC.

Such data will help both prospective postdocs and grad students to “shop” for institutions whose fellows had good career outcomes, says Howard Garrison, public-affairs director of the Federation of American Societies for Experimental Biology. And having such outcomes publicly available could prompt research institutions to be more engaged in career development, because they indicate how well the institutions prepare postdocs for careers.

If the NPA's track record is any indication, the next two years should see more tracking data generated by the increasing numbers of postdoc offices and associations. Hopefully, these data will send both postdocs and the NPA to the top of the class.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik

European Head Office, London
The Macmillan Building,
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
Natureevents: Matt Betteridge (4015)

UK/RoW/Ireland:
Nils Moeller (4953)
Scandinavia/Spain/Portugal:
Claudia Paulsen Young (4973)
France/Switzerland/Belgium/
Italy: Maria Ceraolo (4994)
Germany/Austria/
The Netherlands:
Reya Silao (4970)
Business Development Manager:
Amelie Pequignot (4974)

Advertising Production Manager:
Stephen Russell
To send materials use London
address above.

Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production:
Catherine Alexander

European Satellite Office
Germany: Patrick Phelan
Tel: +49 89 54 90 57 11
Fax: +49 89 54 90 57 20
e-mail: p.phelan@nature.com

US Head Office, New York
75 Varick Street,
9th Floor, New York,

NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building,
2-37 Ichigayatamachi,
Shinjuku-ku,
Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746
Asia-Pacific Sales Manager: Ayako Watanabe
e-mail: a.watanabe@natureasia.com

Talk about toxic

They arrive from other disciplines; they spread into distant fields. Toxicology is a voyage of discovery for scientists with diverse skills, including those of communication. **Ricki Lewis** gets them to open up about it.

Toxicology dominates headlines in the wake of natural disasters, terror events, new illnesses and bizarre side effects from drugs. But the most exciting part, say its practitioners, is that one never knows where an investigation will lead — a toxicologist's turf can be anywhere. Recent studies have traced deaths of Asian vultures to a veterinary drug, identified an inhalation route for the toxin that causes paralytic shellfish poisoning and tracked a rare cancer in rural Turkey to a volcanic building material.

Toxicology considers the effects on organisms of exposure to chemicals, biological agents and radiation. Emphasis on human health extends the field's scope beyond science.

"It is at the interface of science and public policies that many, and often heated, political issues come into play," says William Greenlee, president and chief executive of CIIT Centers for Health Research in Research Triangle Park, North Carolina, a private, nonprofit institute.

Despite toxicology's diverse worksites, the field is cemented by shared problem-solving strategies and the common language of risk assessment. The links are summed up by Kenneth Ramos, director of the centre for genetics and molecular medicine at the University of Louisville Health Sciences Center in Kentucky. "What binds toxicologists together is focus on the study of toxic responses caused by chemical, physical or biological agents; the consequences of toxicity in biological

Making it clear: Ellen Silbergeld is best known for explaining the work of the Environmental Defense Fund to the media.



systems; and measures to prevent or ameliorate the devastating personal and societal consequences of toxicities," he says.

Because toxicology is quite a new, interdisciplinary applied science, it can be approached by different career pathways. So numerous and eclectic are job opportunities that it isn't uncommon for a seasoned scientist to have spent time in academia, industry and government, and finish with private consulting.

Toxicology came into its own in the 1960s. "There was a big awakening about the environment and chemicals," recalls Jay Gandolfi, assistant dean of research and graduate studies in the College of Pharmacology at the University of Arizona in Tucson.

"There has been a real evolution in safety assessment from 40 years ago, when we used a 'dose 'em and count 'em' approach," says Linda Birnbaum, director of the experimental toxicology laboratory at the Environmental Protection Agency (EPA) in Research Triangle Park, North Carolina. "Now analyses are much more molecular, targeted and mechanistic."

The world as laboratory

A career in toxicology might take a scientist to a contaminated well, a crime scene, a courtroom, an analytical chemistry lab or a political hearing. At the Environment Agency in Britain, for example, some scientists pursue basic research at a national centre and others become field-based environment-protection officers or join a policy team. Requirements range from broad training in biology or chemistry to specific expertise, such as in environmental oestrogens.

The International Neurotoxicology Association divides career opportunities into regulatory,

commercial, scientific and medical, says its president David Ray, leader of the MRC Applied Neurosciences Group at the University of Nottingham, UK. Regulatory neurotoxicologists work for national or international agencies. The commercial track includes the drug, bulk chemical and pesticide industries.

All nations face toxicological problems, says Ray, but priorities vary. "In most developed countries, concerns centre on evaluation of new products, relatively low-level toxicity and reduction of accidental exposures. In developing countries perspectives can be very different," he says, citing the use of pesticides in China and Malaysia for suicide. Countries may present specific challenges, he adds, such as the neurotoxin domoic acid, sometimes found in Canadian mussels.

An unusual skill mix

Even though you can now gain a PhD in toxicology, broad training is still valuable. A neuroscience degree can segue into neurotoxicology, or an immunology degree into immunotoxicology. Now a consultant, Skip Matthews of Hertford, North Carolina, merged a background in entomology and an interest in chemistry into a three-decade career in insecticide toxicology at the National Institute of Environmental Health Sciences (NIEHS) in Research Triangle Park.

Ellen Silbergeld, professor of environmental health sciences at the Johns Hopkins University Bloomberg School of Public Health in Baltimore, Maryland, recommends "a strong base in a relevant discipline, either in the biological or chemical sciences or in engineering work". After an undergraduate degree in history and a doctorate in engineering, Silbergeld learned toxicology as a postdoc. She is probably best known for her clear statements to the media, as senior consultant toxicologist for the Environmental Defense Fund in Washington DC.

Communication skills are vital when testifying. Forensic toxicologist Marc Pelletier, of the Centre of Forensic Sciences in Ontario, Canada, faces a courtroom nearly every week. He analyses body fluids for drugs, lectures to students and police officers, sees pathologists and coroners, and, like all toxicologists, writes many reports. After a doctorate in behavioural neuroscience, his training came mostly on the job. "All scientists are required to complete a comprehensive training programme specific to forensic toxicology at the centre, of approximately three years," says Pelletier.

Because most toxicologists work in groups, interpersonal skills are important. "Successful toxicologists are good observers and have learned management by following role models and then customizing the style to fit their situations," says Gandolfi.

Private consulting offers its own challenges, such as discussing risks with clients who aren't science-savvy. "A comfort level in working on a broad range of topics and good communication and on-your-feet problem-solving abilities are useful," says Barbara Beck, a toxicologist at Gradient in Cambridge, Massachusetts.

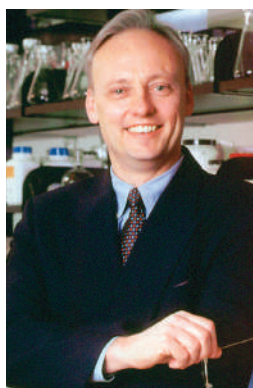
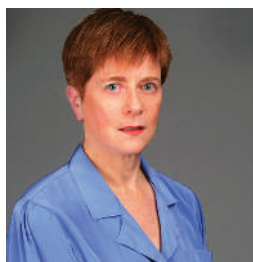
Combining scientific expertise with these talents is a tall order. Rick Schnellmann, chair of pharmaceutical sciences at the Medical University of South Carolina, Charleston, sums it up. "Employers want diverse laboratory skills, the ability to focus on a specific topic as well as looking at the big picture, excellent writing and communication skills, and motivation," he says.

You may want to try a taster before starting a



Keeping a grip: William Greenlee warns of heated political issues.

Tips for toxicologists: you need on-your-feet problem-solving skills, says Barbara Beck; Rick Schnellmann finds that employers want people who can focus on a topic without losing sight of the big picture.



doctorate. The Health Effects Institute in Boston, Massachusetts, a nonprofit organization informing government and industry about air pollution, seeks research assistants with degrees in toxicology, chemistry, epidemiology, biology, mechanical engineering or environmental science. The National Research Centre for Environmental Toxicology in Queensland, Australia, seeks students to help develop a system to remove arsenic from water supplies.

Because doctorates in toxicology are fairly new, many toxicologists entered "through the back door," says Jack Dean, retired president of US science and medical affairs at Sanofi-Aventis in Paris. Dean, whose graduate work was in molecular biology and biochemistry with a focus on immunology, also started the NIEHS immunotoxicology programme.

Pharmacology was the route for many, including James Bond, editor of *Chemico-Biological Interactions* in Durham, North Carolina. "Pharmacology deals with efficacy and toxicology deals with adverse effects," he says. "I became increasingly interested in how chemicals exerted their toxic effects on biological systems."

MaryJane Selgrade, chief of immunotoxicology at the EPA, learned by herself. With a PhD in medical microbiology, she moved into toxicology when she discovered that certain viruses enhance the toxicity of some compounds, rather than the compounds altering immunity to the viruses. "My career and the speciality of immunotoxicology grew up together," she recalls.

Jane Allen, too, taught herself, after training in microbiology and genetics. "I bought textbooks, read journals and attended meetings and continuing-education courses," says Allen, former director of toxicology for GlaxoSmithKline, who is now a consultant in Raleigh, North Carolina.

Toxicologists can maintain their skills with the help of continuing education programmes offered by such organizations as Eurotox, a European federation of individuals and societies, and IUTOX, the international union of toxicology, based in Reston, Virginia.

In the real world, people rarely encounter toxins in isolation. "We experience a suite of chemicals against a backdrop of different genetic backgrounds, drug exposures and foods — we need a combinatorial approach," says Birnbaum. So toxicology is moving away from analysing chemicals one at a time. A study in Salinas Valley, California, for example, probes the susceptibilities of Latino people to several pesticides in their environment. Beck is using a systems approach to analyse residual material on a medical implant.

Rory Conolly, a senior research biologist at the EPA, has been developing computer models of how multiple toxins affect the body since the early 1980s. "Because of the availability of deep data, people are realizing that the only way to deal with the avalanche is to be more sophisticated in how we analyse it — how to store, organize and apply advanced statistics to look for patterns in data," he says.

Toxicologists are such an enthusiastic bunch that many are stymied when asked to name a favourite investigation. Ramos's response is typical: "I have enjoyed every project I have ever worked on!" And with new tools to evaluate multi-layered data, that enthusiasm is certain to grow. "It is an exciting field with the potential for a lot of challenges over the next five to ten years," says Birnbaum.

Ricki Lewis is a freelance writer in Scotia, New York.

MOVERS

Stephen Forrest, vice-president for research and William Gould Dow professor in electrical engineering, University of Michigan, Ann Arbor



1992–2006 James S. McDonnell professor of electrical engineering, Princeton University, Princeton, New Jersey
1997–2001 Chairman, Department of Electrical Engineering, Princeton
1992–1997 Director, Center for Photonic and Optoelectronic Materials, Princeton

As an undergraduate, Stephen Forrest threw a dart at a map to determine the site of his doctorate. It guided the then University of California, Berkeley, physics undergraduate to the University of Michigan to get his PhD studying magnetism. But it was his next major career decision — to accept a position at Bell Labs in Murray Hill, New Jersey, on the development end of the R&D spectrum — that steered him towards the intersection of academia and industry.

Working in development gave Forrest practical training in systems design and economics: “I was becoming an engineer without my knowing it,” he says. At Bell Labs he developed the first practical indium gallium arsenide detector, still used in fibre-optic systems today. After six years, however, he started to get restless for academia and found a natural fit at the electrophysics department at the University of Southern California. There, he became director of the National Center for Integrated Photonic Technology, a research consortium between five universities including the Massachusetts Institute of Technology and Princeton, funded by the Defense Advanced Research Projects Agency.

But the pull of industry would return. Forrest met Greg Olson, then a friendly competitor at RCA labs. Olson — now best known for being the third space tourist — approached Forrest to start a company called EpiTxx. Uninterested in business management, Forrest agreed, instead, to consult. “I’ve had opportunities to become CEOs and presidents of companies, but never considered giving up the freedom of thought in academia,” he says.

Olson wanted Forrest closer to his New Jersey-based operations and so he took it upon himself to submit Forrest’s resumé for the directorship of Princeton University’s Center for Photonics and Optoelectronic Materials (POEM). Building on POEM’s charter to work with industry, Forrest crafted the first agreements allowing company engineers to work at POEM facilities.

Now he’s returning to Michigan in a new capacity — vice-president for research. He is already working on significant initiatives, most notably in energy, aiming to foster more academic-industry collaborations there.

Forrest credits the calibre of his collaborations, notably with Olson and chemist Mark Thompson, as the secret to his success. “You have to bring together skills in today’s world,” he says. “Take the time to teach each other.” ■
Virginia Gewin

RECRUITERS & ACADEMIA

What makes a good PhD student?

Doing a PhD should be fun and rewarding, because you can spend all your working time discovering things and pursuing ideas — and getting paid for it, without any administrative responsibilities. Those who stick with a career in science do so because, despite the relatively poor pay, long hours and lack of security, it is all we want to do.

Unfortunately most new PhD students are ill-prepared, and as a consequence very few will fulfil their aspirations to be independent scientists. The main reasons for this are the ‘grade creep’ inherent at most universities, making it difficult to identify the really talented first-class graduates from the rest, and the pressure on universities to graduate as many PhD students as possible. The consequence is that we enrol far too many of them without telling them clearly what doing a doctorate should entail. We therefore set ourselves, and the students, on a path of frustration and disappointment.

So what should we be telling prospective PhD students?

- Choose a supervisor whose work you admire and who is well supported by grants and departmental infrastructure.
- Take responsibility for your project.
- Work hard — long days all week and part of most weekends. If research is your passion this should be easy, and if

it isn’t, you are probably in the wrong field. Note who goes home with a full briefcase to work on at the end of the day. This is a cause of success, not a consequence.

- Take some weekends off, and decent holidays, so you don’t burn out.
- Read the literature in your immediate area, both current and past, and around it. You can’t possibly make an original contribution to the literature unless you know what is already there.
- Plan your days and weeks carefully to dovetail experiments so that you have a minimum amount of downtime.
- Keep a good lab book and write it up every day.
- Be creative. Think about what you are doing and why, and look for better ways to go. Don’t see your PhD as just a road map laid out by your supervisor.
- Develop good writing skills: they will make your scientific career immeasurably easier.
- To be successful you must be at least four of the following: smart, motivated, creative, hard-working, skilful and lucky. You can’t depend on luck, so you had better focus on the others! ■

Georgia Chenevix-Trench is principal research fellow at the Queensland Institute of Medical Research, Royal Brisbane Hospital, Herston, Australia.

▶ www.qimr.edu.au/research/labs/georgiat/Guideforphds.doc

GRADUATE JOURNAL

Valuable diversions

Passion for science can make it hard to stop thinking about work outside office hours. One solution is to engage in hobbies that force you to switch off from thoughts of work, such as sports, crafts or, in my case, music. At choir rehearsals, I effortlessly shift my focus off small black ants onto the small black notes on the sheet music.

Choirs, teams or dance groups can offer a welcome break from the scientific world, where social gatherings inevitably lead to talking shop. In the choir, I meet people in different professions and life situations. While I may tell the occasional ant story, pub discussions are as likely to revolve around life as a casino dealer, teacher or full-time mother.

But perhaps most importantly, performing in a concert or learning a new piece of music gives me a feeling of instant gratification that research sometimes lacks. The average biological project can take years from planning to writing up, and when your inbox finally delivers that longed-for letter of acceptance, you’re already deep into the next project. Short-term successes outside science — learning a new language, competing in a race or performing in a play — can boost your morale, giving you strength to continue in the marathon run towards graduation. ■

Katja Bargum is a graduate student in the Department of Biological and Environmental Sciences at the University of Helsinki, Finland.

The aching of Dion Harper

A case of phantom body syndrome.

Arthur Chrenkoff

After the man with steel-blue eyes had killed me, he knelt down next to my body, pulled out a thin-bladed knife from a sheath strapped to his shin, and with one quick movement cut off my right ear. That wasn't a part of his contract; I guess he must have had a fetish for trophies.

Only moments earlier I had been pleading with him for my life, cornered like a rat, my back pressed against a locked exit door at the lowest level of the Metro Tower underground hoverpark. "I'm not Dion," I cried through chattering teeth, the realization of the imminent end squeezing the air out of my windpipe. "For God's sake, I'm not Dion. I don't..."

"Ain't that what all of them always say," the man with steel-blue eyes cut me off, a tinge of veteran's weariness colouring his voice. Behind us lay the trail of death — his dead associates, killed by my dead bodyguards — until there was only me, him and his gun remaining, all claustrophobically entombed underneath two hundred storeys of iron, polymer and glass.

"Either way, Dion, we'll find out soon enough," my killer said and then he extended his arm, aiming at my torso.

I have now watched my last moments through my killer's eyes. Minutes before the man with steel-blue eyes was dispatched with a single bullet to the back of the head, one of my best lieutenants and his support team downloaded a cache of my killer's memories. It took an outfit I subcontract in West Virginia almost 24 hours of non-stop work to separate a few useful grains from the chaff of the man's — by the way, his name was Heikke — life memories.

And so, a few hours later, I sat on a synth-skin sofa with a glass of scotch in my hand, watching the misty holo replay of the most interesting tidbits once stored in Heikke's mind; his contacts with mid-level operatives of the Syndicate, his pursuit and killing of me, and finally his own capture. For some reason, the guys in West Virginia thought that the bondage session Heikke had engaged in with a prostitute in Vegas the night before he killed me was somehow relevant. I fast-forwarded

through it. Sometimes, I think, they do these things as a bit of a joke.

The rain is falling hard now, streaking the window and obscuring the view of the meadow as it gently slopes for a few hundred yards towards a little brook. One of Bruckner's symphonies is playing in the background, masking the gentle buzz of information streaming into five terminals around the room, connecting me to my far-flung empire.

I'm aching now.

inside me compounds, growing like tree rings around my self, layer upon layer of longing. But it's still bearable. And, in the end, better than me, isn't that right? So I'll deal with the ache.

There are another seven Dions waiting on stand-by. Now that the man with steel-blue eyes has eliminated my previous public self, one of the reserve ones is being readied to enter circulation.

It's not cheap, these lives of mine. As the saying goes, only multibillionaires can afford to be multi. It costs ten million

Standard Units to successfully clone an individual, another ten or more to bring the clone up, double that when you go for accelerated growth. And then there are the optional extras like the fake memory implants, which nowadays aren't all that optional anymore. After all, you can't make

it too easy for those who are hunting you. The Dion one before last was so good and so believable that for about 16 months the Syndicate really believed that this time they had finally managed to kill me. This hiatus in our dirty little tit-for-tat has bought me some very valuable time to further rearrange my many affairs.

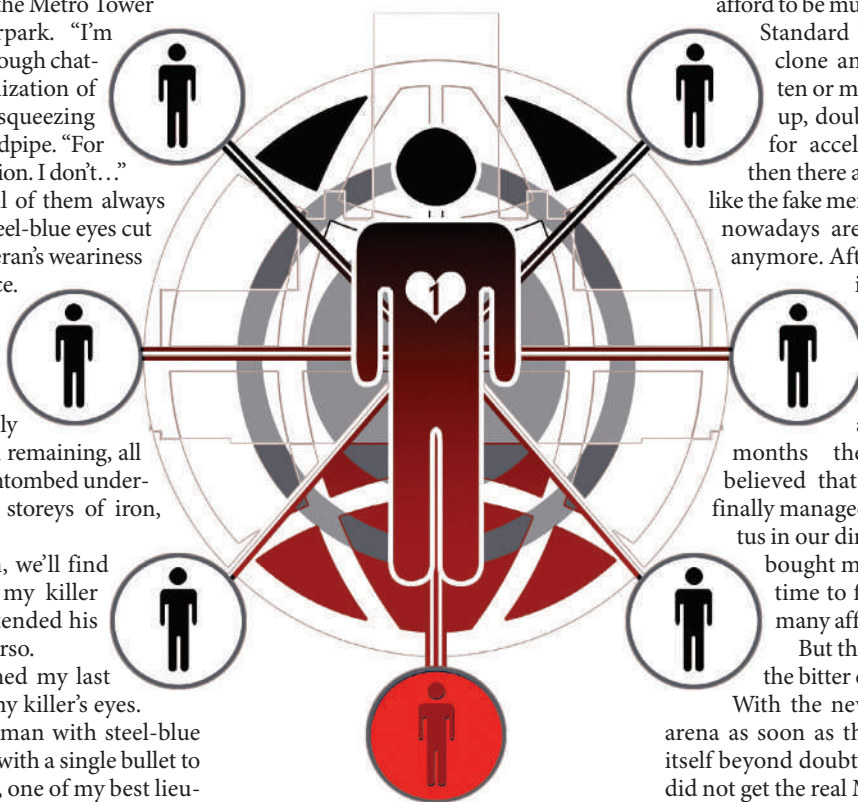
But the war will go on, until the bitter end — theirs or mine.

With the new Dion entering the arena as soon as the Syndicate satisfies itself beyond doubt that once again they did not get the real McCoy, I'll still have a half-a-dozen of me kept in reserve, ready to be called in at a short notice. And I can and will have still more. The sky's the limit. Or, more precisely, the limit is my resources, and they are considerable.

But there is only one me — the real me. Many vessels, though filled with the breath of life, but only one soul to go between them all. And I'm its proud sole owner. I'll be damned if I will let the Syndicate, or anyone else, steal that most precious of all my possessions.

I am Dion Harper. The one — but not only.

Arthur Chrenkoff is a Polish-Australian retired blogger. His first novel, *The Night Trains* is being published by Cold Spring Press. To the best of his knowledge, he hasn't been cloned yet.



JACEY